THE HANDBOOK OF POST CRISIS FINANCIAL MODELING

EDITED BY
EMMANUEL HAVEN
PHILIP MOLYNEUX
JOHN O.S. WILSON
SERGEI FEDOTOV
MERYEM DUYGUN

The Handbook of Post Crisis Financial Modeling

The Handbook of Post Crisis Financial Modeling

Edited by

Emmanuel Haven
Professor, School of Management and Institute of Finance, University of Leicester, UK

Philip Molyneux
Dean of College of Business, Law, Education and Social Science, Bangor University, UK

John O.S. Wilson
Professor of Banking and Finance, University of St Andrews, UK

Sergei Fedotov
Professor of Applied Mathematics, University of Manchester, UK

and

Meryem Duygun
Professor of Banking and Finance, Hull University Business School, UK

palgrave
macmillan

First published 2016 by
PALGRAVE MACMILLAN

Palgrave Macmillan in the UK is an imprint of Macmillan Publishers Limited, registered in England, company number 785998, of Houndmills, Basingstoke, Hampshire RG21 6XS.

Palgrave Macmillan in the US is a division of St Martin's, Press LLC, 175 Fifth Avenue, New York, NY 10010.

Palgrave Macmillan is the global academic imprint of the above companies and has companies and representatives throughout the world.

Palgrave® and Macmillan® are registered trademarks in the United States, the United Kingdom, Europe and other countries.

ISBN 978-1-349-57578-7 ISBN 978-1-137-49449-8 (eBook)
DOI 10.1007/978-1-137-49449-8

This book is printed on paper suitable for recycling and made from fully managed and sustained forest sources. Logging, pulping and manufacturing processes are expected to conform to the environmental regulations of the country of origin.

A catalogue record for this book is available from the British Library.

Library of Congress Cataloging-in-Publication Data

The handbook of post crisis financial modeling / edited by Emmanuel Haven, Professor, School of Management and Institute of Finance, University of Leicester, UK, Philip Molyneux, Dean of College of Business, Law, Education and Social Science, Bangor University, UK, John O.S. Wilson, Professor of Banking and Finance, University of St Andrews, UK, Sergei Fedotov, Professor of Applied Mathematics, University of Manchester, UK, Meryem Duygun, Professor of Banking and Finance, Hull University Business School, UK.

pages cm

1. Economics, Mathematical. 2. Finance – Mathematical models.
3. Global Financial Crisis, 2008-2009. I. Haven, Emmanuel, 1965- editor.

HB135.H3576 2015
332.01′5195–dc23 2015029078

To Sophie, Nath and Sam
Emmanuel Haven

To Delyth, Alun, Catrin, Gareth, Gethin, Lois and Rhiannon
Phil Molyneux

For my nephew Oliver
John O.S. Wilson

To Masha, Petja and Tema
Sergei Fedotov

Contents

List of Figures

List of Tables

Preface and Acknowledgments

After the onset of the 2008 financial crisis, many media reports questioned whether the academic finance community should have been able to predict this crisis. It is certainly fair to say that prior to 2008 very few contributions to the empirical finance literature modeled events as extreme as those which occurred during the crisis. However, following the crisis there is now more interest in extreme event modeling than at any other time in history. In addition to increased interest in extreme event modeling, many facets in mainstream finance have evolved since the crisis. For example, the issue of using liquidity as an explicit input in mathematical and empirical based modeling has become prominent.

The contributors to the Handbook comprise experts from a variety of fields, such as financial history, banking and mathematical finance. The content of the Handbook provides a blend of theoretical, empirical, policy and practitioner insights of many of the developments that have taken place since the onset of the financial crisis in 2008.

What better place to start our journey than with history? History is replete with dramatic financial events, and while it is debatable that any of those events have really been as extreme as what happened in 2008, there is still an array of useful information embedded in prior financial experiences. Historiography is the topic covered in the first chapter of this Handbook. The contribution of Peter L. Rousseau outlines the evolution of the US financial system through the colonial period highlighting the major events and economic challenges posed to the US economy. The chapter shows that an important lesson can be learned from studying history: namely that the presence of a strong central bank acting as lender of last resort is of paramount importance for maintaining economic stability.

Following this historical foundation, the Handbook proceeds on a "technical" journey, with a group of chapters that revolve around studies on the topic of banking. Sanja Jakovljević, Hans Degryse and Steven Ongena review the empirical monetary transmission literature. The authors assert that policy making has been well informed by input from recent empirical studies. However, there are problems in a variety of areas such as overly restrictive credit reporting requirements and the issue of the level of universality of variable definitions without which international comparisons of modeling results become very difficult. The chapters by Rhiannon Sowerbutts/Peter Zimmerman and Ali al-Nowaihi/Sanjit Dhami examine issues related to government responses toward banking following the crisis. Sowerbutt and Zimmerman's chapter explores the important problem of market discipline in banking. It considers the reasons why such discipline can break down and examines whether bank opacity or transparency is socially optimal. In the chapter by al-Nowaihi and Dhami, a model is proposed which shows that there exists an optimal institutional response to the well-known liquidity trap (where bonds and money become substitutes due to zero nominal interest rates). Hulusi Inanoglu, Michael Jacobs, Jr., Junrong Liu and Robin Sickles explore the "too-big-to-fail" controversy. The authors make use of an impressive suite of econometric techniques in attempting to answer the thorny issue of whether "too-big-to-fail" banks are efficient. The authors argue that bailing out such large banks may come at a huge cost (with problems such as moral hazard on one side and weakening levels of international competitiveness on the other). The chapter by Thomas Weyman-Jones addresses another very interesting problem regarding how a measure of efficiency can be used to assess bank recapitalization costs. This work presents an extension to the broad literature on modeling bank efficiency and provides insights into the link between efficiency and bank safety.

A second group of chapters investigates modeling responses to the 2008 crisis from a variety of angles, other than the banking perspective. The chapter written by Mark H. A. Davis explores how one can quantify the level of uncertainty in option pricing when, using the words of Davis, it is "occasioned by the variety of models used." His chapter highlights an important issue most of us face when we model, namely, the extent to which "minimal assumptions" are used. A key ingredient which appears by virtue of necessity in many financial models is the concept of information. The chapter by Jérôme Detemple and Marcel Rindisbacher defines the concept of Private Information Price of Risk (PIPR) as representing "the incremental price of risk assessed when private information becomes available." Their work investigates how PIPR behaves in both discrete and continuous time models. The authors show that PIPR really quantifies the information content of the signal. The authors also mention the difficulties involved in modeling pricing relationships when some private information converts into public information. The chapter by Igor Evstigneev, Thorsten Hens and Klaus Reiner Schenk-Hoppé considers an alternative approach to traditional equilibrium analysis in finance. One of the central arguments in this new approach is that market dynamics are a sequence of consecutively related short-run equilibria. Jukka Isohätälä, Nataliya Klimenko and Alistair Milne consider a new approach to macroeconomic modeling. Their approach uses non-linear continuous time specifications of economic dynamics.

Overall, the work contained in the aforementioned four chapters has relevance to the modeling implications which arise as a consequence of the 2008 financial crisis. In effect, all of the four chapters provide excellent frameworks that can be used to calibrate essential economic and financial variables that are used to model crises. This Handbook is rounded off with a last (but surely not least important) chapter on using a resolutely different approach to modeling in finance. In the chapter by Fabio Bagarello, a simplified stock market is proposed which allows for the definition of a so-called Hamiltonian on this simple market. The Hamiltonian is an operator, and so are the operators which trigger share price changes and market supply changes. This chapter sits within the wider literature that applies concepts of physics to economics and finance.

This Handbook would not have been possible without the financial support of the ESRC (Economic and Social Research Council) Seminar Series. The seminar series upon which this Handbook is based was entitled "Financial Modeling Post-2008: Where Next?" and since the ESRC is the largest funding body in the United Kingdom which funds research in the social sciences we, as editors, were quite thrilled to receive funding from this body. The financial assistance received from this source allowed us to organize four different seminars held first at Bangor University, then the University of St. Andrews, followed by the University of Manchester and the University of Leicester. The Handbook contains some of the papers presented at those seminars. None of those seminars would have been possible without the excellent administrative support of Karen Williams (Bangor University), Shona Deigman (University of St. Andrews), Steven Falconer (University of Manchester) and Daksha Patel (University of Leicester).

Finally, the support provided by staff at the ESRC also needs to be stressed. Claire Mussen was always available to answer urgent (and important) queries. The editors also want to thank the contributors to this Handbook and all the speakers who presented work at the four ESRC seminar series. We also want to thank Aimee Dibbens and Grace Jackson (both at Palgrave Macmillan Publishers) for their very able and valuable assistance in making this Handbook a reality.

Notes on Contributors

Fabio Bagarello is a professor in the Department of Energy, Information Engineering and Mathematical Models at the University of Palermo, Italy. He is the author of over 160 journal articles and of several books on quantum tools for macroscopic systems, functional analysis, quantum mechanics with non self-adjoint operators and classical mechanics.

Mark H. A. Davis has spent most of his career at Imperial College London, which he joined in 1971 on completion of his PhD at the University of California Berkeley, working on modeling and control of stochastic systems. In 1995 he was hired by the London-based investment bank Mitsubishi Finance to run a front office quant group providing pricing models and risk analysis for fixed-income, equity and credit-related products. He returned to Imperial in 2000 to start a graduate program in Mathematical Finance. His work has been published widely, recent book being *Risk-Sensitive Investment Management* (2014, with Sébastien Lleo).

Hans Degryse is Professor of Finance in the Department of Accountancy, Finance and Insurance of the KU Leuven. He is a research fellow at the CEPR, CESIfo, the European Banking Center (EBC), and TILEC, and is a member of the academic council of EBC. Before joining Leuven in 2012, he was Professor of Finance at Tilburg University, Netherlands. His research focuses on financial intermediation, including theoretical and empirical banking as well as market microstructure. His articles have appeared in many journals including the *American Economic Review, Journal of Finance, Journal of Financial Economics, Review of Financial Studies, Management Science, Journal of Financial Intermediation,* and the *Economic Journal*, and he has presented in leading international conferences such as the American Finance Association, the Western Finance Association, the European Finance Association, and the Financial Intermediation Research Society. He co-authored, with Moshe Kim and Steven Ongena, the graduate textbook *Microeconometrics of Banking: Methods, Applications and Results* (Oxford University Press). He is currently an associate editor of the *International Review of Finance* and the *Review of Finance*.

Jérôme Detemple is Professor of Finance and Everett W. Lord Distinguished Faculty Scholar at Boston University Questrom School of Business. He holds a PhD in Finance from Wharton, a Doctorat d'État ès Sciences Économiques from Université de Strasbourg, as well as degrees from ESSEC and Université de Paris-Dauphine. He is the author of over 50 articles that have appeared in journals, including *Econometrica, Journal of Finance, Review of Financial Studies* and *Mathematical Finance*. He is the current Editor-in-Chief of *Mathematical Finance* and an associate editor of the *Journal of Financial Engineering*. His research focuses on portfolio selection, asset pricing and derivative securities.

Sanjit Dhami is Professor of Economics at the University of Leicester, UK. He holds an MA from the University of Toronto and an MPhil from the Delhi School of Economics. He received the Chancellor's gold medal in his undergraduate and postgraduate degrees. He has previously taught at the Universities of Toronto, Essex and Newcastle-Upon-Tyne. His articles have appeared in the *Journal of Public Economics, European Economic Review, Journal of Mathematical Psychology, Oxford Economic Papers,* and the *Journal of Economic Behavior and Organization*, among others. His main area of research is behavioral economics, working closely with his colleague Ali al-Nowaihi.

Igor Evstigneev is Professor of Mathematical Economics at the University of Manchester, UK. A mathematician by training, he has published more than 150 papers in top journals in mathematics, economics and finance. His research interests include stochastic dynamic models in economics and finance, mathematical behavioral finance and stochastic dynamic games. He served as associate editor for *Economic Theory*, *Journal of Mathematical Economics*, and *Random Operators & Stochastic Equations*. During the past decade, he has worked primarily in the field of mathematical financial economics, on which he has recently completed a book (with Hens and Schenk-Hoppé).

Thorsten Hens is Swiss Finance Institute Professor of Financial Economics at the University of Zürich, Switzerland and Adjunct Professor of Finance at the Norwegian School of Economics, Norway. He studied at Bonn and Paris and held positions in Stanford and Bielefeld. His main research areas are behavioral and evolutionary finance. In his behavioral finance research he develops tools to avoid investment mistakes. In his evolutionary finance research he studies the interaction of investment strategies in order to determine properties for long run survival. He has written five books and more than 50 papers that have appeared in top peer reviewed journals.

Hulusi Inanoglu is a senior economist in the Quantitative Risk section of Banking Supervision and Regulation at the Federal Reserve Board. Inanoglu co-leads the Federal Reserve Trading Book Qualification Team which advises senior management on banks' risk models used for market risk, and counterparty credit risk regulatory capital calculations. Before joining the Federal Reserve Board in 2009, he worked at the OCC, a bureau of the U.S. Department of the Treasury which charters, regulates and supervises national banks. He has a BS in Mechanical Engineering from Middle East Technical University, an MA in Economics from the University of Colorado and a PhD in Economics from Rice University.

Jukka Isohätälä is a Theoretical Physics PhD graduate (Condensed Matter Theory) from the University of Oulu, Finland. His areas of interest include macrofinance modeling, computational economics, and interdisciplinary approaches. In his post-doctoral research, he has worked on macroeconomics projects at Loughborough University's School of Business and Economics and in the Department of Physics.

Michael Jacobs, Jr., PhD, CFA, is a Director in PwC's Risk Advisory practice in risk and regulation, a leader in the models and methodologies development and validation practice. He has 20 years of experience in the risk modeling practice across various risk types (credit, market, enterprise and operational), leading projects in model development and validation, supervisory risk modeling issues, model risk management, economic capital modeling, capital adequacy analysis, stress testing, loss forecasting, credit risk of wholesale loan portfolios, and optimizing portfolio risk measurement and management through the use of advanced econometric and computational statistical modeling approaches.

Sanja Jakovljević is a PhD student in Finance at KU Leuven, Belgium (on leave from the Croatian National Bank). She holds an MSc in Macroeconomics from Tilburg University, and she graduated from the University of Zagreb with an MA in Statistics and a BSc in Macroeconomics. Prior to her Master's, she worked in the Research Department of the Croatian National Bank, and she is currently associated with the Research Department of the National Bank of Belgium. Her research interests include banking and regulation, macroeconomics, and linkages between finance and innovation.

Nataliya Klimenko is a post-doctoral researcher at the University of Zürich, Switzerland. She holds a PhD in Finance from Aix-Marseille School of Economics. Her research interests are in the areas of bank regulation, corporate finance, macro finance and dynamic contract theory.

Junrong Liu specializes in credit risk modeling, mortgage default and prepayment transition modeling, loss severity modeling, mortgage origination forecasting, housing price and interest rate stochastic processes and dynamic Monte Carlo simulations. She also has experience with data mining techniques, including multivariate regression/classification, logistic regression and regression tree methods. She holds an MA in Statistics and a PhD in Economics from Rice University.

Alistair Milne is Professor of Financial Economics in the School of Business and Economics, Loughborough University, UK and was previously with Cass Business School, the Bank of England, the University of Surrey, London Business School, HM Treasury and the Government of Malawi. He is the author of a comprehensive account of the global credit crisis, *The Fall of the House of Credit* (2009) and holds a PhD in Economics from the London School of Economics. He works on a range of issues of relevance to central banks and policy makers, including banking regulation, monetary transmission and financial infrastructure and technology.

Ali al-Nowaihi is Professor of Economics at the University of Leicester, UK. He holds an MSc in Mathematics from the University of London and a PhD in Oligopoly Theory from South Bank University, London. His articles have appeared in the *Journal of Algebra, Journal of Economic Theory, Journal of Monetary Economics, European Economic Review, Cognitive Psychology, Journal of Mathematical Psychology* and the *Journal of Economic Behavior and Organization*, among others. Since 2000 his main area of research has been in behavioral economics, working closely with his colleague Sanjit Dhami.

Steven Ongena is Professor of Banking at the University of Zürich, Switzerland and the Swiss Finance Institute. He is also a research professor at Bangor University and a research fellow of CEPR. He is the author of more than 45 papers that have appeared in refereed academic journals, including the *American Economic Review, Econometrica, Journal of Finance, Journal of Financial Economics, Journal of International Economics,* and *Review of Finance*, among other journals, and he has written more than 45 papers that have appeared as chapters in books and other collections. He is currently an associate editor of the *Journal of Finance*, a co-editor of the *International Review of Finance*, and he serves as an associate editor for a number of other journals. In 2009 he received a Duisenberg Fellowship from the European Central Bank, and in 2012, a NYU Stern-Fordham-RPI Rising Star in Finance Award.

Marcel Rindisbacher is Associate Professor of Finance at Boston University Questrom School of Business, USA. He holds a PhD in Economics from Université de Montréal, an MSc in Mathematical Economics and Econometrics from the London School of Economics and a BA in Economics from the Universität Bern. His research in financial economics and econometrics has been published in the *Journal of Finance, Review of Financial Studies, Journal of Financial Economics, Mathematical Finance, Finance and Stochastics* and *Journal of Econometrics*. He is an associate editor of *Mathematical Finance*. His research focuses on portfolio selection, asset pricing and computational finance.

Peter L. Rousseau is Professor of Economics at Vanderbilt University, Tennessee, USA and Secretary-Treasurer of the American Economic Association. A macroeconomist and economic historian who studies the role of financial markets and institutions in growth and development, he is particularly interested in the monetary history of the United States and Europe and has published widely in this area. His work also considers how financial markets assist in spreading transformative technological changes through an economy.

Klaus Reiner Schenk-Hoppé is Professor of Financial Economics at the University of Manchester, UK and Adjunct Professor of Finance at the Norwegian School of Economics, Norway. He held the Centenary Chair in Financial Mathematics at the University of Leeds, 2005–2014, and was an associate professor at the Institute of Economics, University of Copenhagen, 2002–2004. Klaus has published extensively on economics, finance and mathematics, edited several special issues, and is co-editor of the *Handbook of Financial Markets: Dynamics and Evolution* (2009, North Holland). His research interests comprise computational and behavioral evolutionary models of financial markets, mathematical financial economics, and dynamic and mathematical economic theory.

Robin Sickles is the Reginald Henry Hargrove Professor of Economics and Professor of Statistics at Rice University, USA. He holds a PhD from the University of North Carolina and a BS in Economics from Georgia Institute of Technology. His research focuses on productivity and empirical industrial organization, panel data econometrics, and nonparametric econometrics. He has taught or has been a visiting scholar at the University of Pennsylvania, George Washington University, University of North Carolina at Chapel Hill, University of Michigan, The European Institute of Business Administration (INSEAD), C.O.R.E. and the Institut de Statistique-Université Catholique de Louvain, Wissenschaftzentrum Berlin (WZB), the Hausdorff Research Institute for Mathematics, University of Bonn, University of Queensland, the University of Melbourne, and the University of Loughborough.

Rhiannon Sowerbutts is a senior economist in the Financial Stability Strategy and Risk Directorate of the Bank of England. Her work focuses on macroprudential policy, capital flows and international banks. She holds a PhD and an MSc in Economics from Universitat Pompeu Fabra and a BSc in Economics from the London School of Economics.

Thomas Weyman-Jones is Emeritus Professor of Industrial Economics in the School of Business and Economics at Loughborough University, UK and an associate editor of the *Journal of Productivity Analysis*. He has a general research interest in measuring the impact of regulation and competition, and particular interests in efficiency and productivity analysis, using the techniques of stochastic frontier analysis, stochastic non-parametric approach to efficiency analysis and data envelopment analysis. His recent work includes applications to banking systems and in the energy industries.

Peter Zimmerman is a senior economist at the Bank of England, which he joined in 2007. His research interests include banking regulation, market discipline, network theory, shadow banking and macroprudential policy. Peter holds a BA in Mathematics from the University of Cambridge and an MSc in Economics from Trinity College Dublin. Currently he is studying for a PhD in Financial Economics at the University of Oxford.

1
Financial Development and Financial Crises: Lessons from the Early United States

Peter L. Rousseau

Introduction

The financial history of the United States is unique in that it includes multiple experiments with currency and banking systems that were accompanied by the rapid emergence of financial markets. This history is also reasonably well documented, and includes relatively frequent disruptions in the form of financial crises. More importantly, the U.S. experience from the colonial period through World War I holds lessons for understanding the interrelated roles of financial development, globalization, and financial crises in economic growth and stability, and these lessons are not entirely remote from the global events of 2007–2009. This chapter highlights some of these lessons through a financial historiography of the period that emphasizes the key role of central banks in economic stability, the dangers of allowing political expediency to drive economic outcomes, and the pitfalls of allowing either to gain excessive influence over financial and monetary policies.

The colonial period, the time surrounding the War of 1812, and the long period from 1836 until the founding of the Federal Reserve in 1914, all saw the United States without a federal bank. Although there were improvements from one episode the next, all three had more than their share of financial crises. In contrast, the two periods prior to the Fed when the United States did have a federal bank (1791–1811 and 1817–1836) saw greater financial stability, though more in the first period than the second. This chapter focuses on links between growth, volatility, the vulnerability to crises across pre-Fed U.S. history, and their implications for gaining a better understanding of the issues surrounding financial modeling in today's post-crisis world.

1 The colonial period

The colonial period of pre-U.S. history commenced as British settlers migrated to North America in the early 17th century and ended with the fledgling nation's Declaration of Independence from England on July 4, 1776. Most of the colonies engaged in some of the world's earliest experiments with paper money. Although paper money had circulated at various times in China during and after the Tang Dynasty (618–907 A.D.), the British North American colonies were the first to use it as a permanent financial instrument. The various colonial legislatures, starting with Massachusetts in 1690, printed this money, called "bills of credit," with the consent of the colony's Governor (i.e., the representative of the British crown) and then used the bills to purchase goods and services. These included payments to troops defending the colonies from threats by French, Spanish, and Native American forces, but some colonies also loaned the bills out to settlers to fund land purchases. The bills were officially backed by only the faith and credit of the issuing colony, but provisions accompanying their issue usually promised redemption at full value in lieu of taxes at pre-specified future dates.

When redeemed on schedule, the monetary theories of Sargent and Wallace (1981) and Sargent and Smith (1987) indicate that agents will increase their holdings of otherwise un-backed paper money in anticipation of a decrease in its supply on each redemption date. These promises of redemption (i.e., backing by anticipated taxes) maintain sufficient demand for the paper money so that new issues lead to a general level of prices that is smaller than the proportional increase specified by the quantity theory of money. The quantity theory states that prices should be in line with the money stock when velocity (V) and transactions (Y) are held constant in the "equation of exchange:" that is,

$$\mathrm{MV} = \mathrm{PY}. \tag{1.1}$$

Indeed, issues of paper money in several of the colonies, such as New York, New Jersey, and Pennsylvania were successful insofar as prices did not advance in the same proportion as the circulation. Since colonial monies are generally believed to have traded at floating exchange rates with each other and with the British pound (McCusker 1978), these arrangements provided the individual colonies with some independent control of their monetary policies.[1] But in cases where currency issues expanded excessively or the legislature failed to burn the bills as promised upon receipt as tax payments, doubts would arise among the public about the eventual redemption of outstanding bills at face value. When this occurred in New England after 1740 and in the Carolinas, the first financial crises in what would become the United States ensued.

South Carolina is a case in point, seeing large injections of currency in the 1710s, in 1730, and again from 1755–1760. In the first two instances the emissions were likely responses to threats, actual and perceived, from neighboring Spanish and Native American forces, while the final inflation coincided with the Seven Years War. Although the emissions may well have been put to good use in defending the colony, all came at the expense of a sharp decline in the value of the bills and severe losses to those left holding them during the fall. In New England, currency issues by Connecticut, Massachusetts, New Hampshire, and Rhode Island tended to pass at par for purchases across colonial boundaries, giving the region have the characteristics of an early currency union. Against this backdrop, the colony of Rhode Island, with one-sixth the population of Massachusetts, emitted quantities of currency that made its amount in per capita terms diverge rapidly, reaching more than five times that of its neighbors by 1840 (Rousseau 2006, 104). When New Hampshire increased its issues in response during the mid-1840s, the implicit taxes imposed by these two states on Connecticut and Massachusetts through depreciation of their own bills caused the system to unravel. By 1751, the British Parliament had passed the "Currency Act," which forbade further issues of bills of credit in New England, in effect placing the region on a specie standard for the remainder of the colonial period.

The worst case of over-issuance of fiat currency, however, came shortly after the United States declared its independence from Britain. On the eve of the Revolutionary War, the national legislature, called the Continental Congress, agreed to issue bills of credit to finance the conflict. These bills, called "continentals," were backed only by vague promises of specie redemption in the future, and presumably only if independence was achieved. Although the continentals allowed the nation to finance the early stages of the war, they began an unmitigated decline in 1879 to reach virtual worthlessness by 1781, and remained there until ultimately redeemed at a rate of 100 continentals to a single dollar when the new unit of account was introduced.

It is a little appreciated fact that the United States, buoyed by its new constitution, began its federal history with a default on obligations to its domestic note holders. Those in favor of the default argued that it was expected, that redemption at full value would primarily benefit speculators who had purchased the bills as option-like instruments that were deeply "out of the money," and that as bearer instruments it would be impossible to identify those individuals who actually lost wealth as the continentals plummeted in the late 1770s. It turned out, however, that the default was essential to the nation restoring its credibility and creditworthiness in the international community.

2 The turnaround

Once it seemed clear that the nation would default on the continentals, questions of whether the federal government should have the right to charter corporations and whether individual states should be permitted to issue paper money came to the forefront late in the process of developing the U.S. Constitution in Philadelphia in 1787. When the final document forbade individual states from issuing currencies and included a right for the federal government to "mint coins and regulate their value" through means "reasonable and proper," however, it was not fully understood that this would imply a privatization of the money creation process. Of course, the outright ban on state currencies and the phrases quoted above created an impression that the federal government would be responsible for providing money, yet the coinciding ban on the federal chartering of corporations rendered the form through which the government would assume this responsibility unclear.

Indeed, Alexander Hamilton, the nation's first Secretary of the Treasury, removed any uncertainty by pressing forward after ratification of the Constitution by the states in 1789 with a proposal to charter a federal corporation – the (First) Bank of the United States. The entity would have an authorized capital of $10 million, with only 20 percent held by the government and the remainder by private investors. The Bank would act as fiscal agent of the federal government, holding its deposits and arranging for disbursements, and would issue its own specie-backed notes to circulate among the public. Some in the early Congress considered the federal charter of any corporation, including a "government" bank, as unconstitutional, and these sentiments persisted through the generation of the founding fathers and into the next. Nevertheless, Hamilton used the "necessary and proper" clause in the Constitution to justify the charter and then steered it through the Federalist Congress. The First BUS started operations in 1791.

Rousseau and Sylla (2005) describe the "Federalist financial revolution" as the set of innovations that brought the Bank into existence and followed on its heels. Hamilton had learned from the Bank of England (founded in 1694) and the Bank of Amsterdam (founded in 1609) how the ability to tender government debt in exchange for shares in a government bank could improve the state of a nation's finances. And improvement was certainly needed given that the federal government and individual states were awash in debt from the War of Independence, with bonds selling at pennies on the dollar in the mid-1780s. Hamilton describes the plan in his 1790 *Report on the Bank*, in which he advocates for a privately managed, limited liability corporation divided into 25,000 transferable equity shares with a par value of $400 each. The *Report* calls for the federal government to purchase its 20 percent of the shares using a loan from the Bank to be repaid in installments over a 10-year period. Private investors would be offered up to 80 percent of the shares, with one fourth payable in specie and three-fourths payable in U.S. bonds paying six percent interest. The "6's," as they were called, represented a restructuring of the federal and various state debts. Even though this innovation had been used successfully a

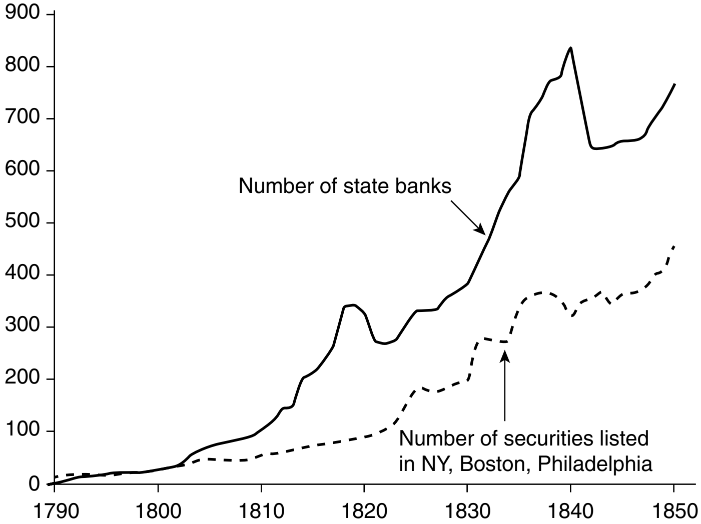


Figure 1.1 The number of state banks nationwide and security listings in three cities, 1790–1850

century earlier in England, it was still something of a surprise that the market value of the new U.S. 6's sprung rapidly to par. By proceeding to make interest payments to foreign and domestic bondholders in hard money payable in the major cities, including London, Hamilton restored the credit standing of the United States and enhanced its ability to draw in capital from abroad. By defaulting on the continentals, Hamilton had made a difficult decision in favor of reliably servicing the restructured debt.

What happened next is nothing short of extraordinary. The number of private banks chartered by individual states rose rapidly to the point where the United States became the most banked nation in the world (Rousseau and Sylla 2005, 5–6). Starting with only three banks in 1790 – one each in New York, Philadelphia, and Boston – the nation attained 31 banks by 1800. Figure 1.1, which shows number of banks from 1790 to 1850, indicates that by 1811, when the 20-year charter of the First BUS was due for renewal, there were 117 banks, and that this expanded to 330 by 1825. Even in England, which had experienced its financial revolution a century earlier, the number of country banks in 1811 stood at only 230. By 1825, Sylla (1998, 93) shows that the United States had roughly $90 million in banking capital, which was 2.4 times the banking capital of England and Wales combined.

But this is moving ahead too far in the account. The innovation of the restructured 6's and the transferability and popularity of shares in the Bank led to the emergence of markets for trading these instruments. Indeed, as Figure 1.1

also shows, the growth in the number of securities traded in the three major cities (New York, Boston, and Philadelphia) from 1790 to 1850 was extraordinary. Starting with a handful of government securities in 1790, by 1825 the United States saw 187 different securities trading in these cities compared to a total of 230 securities trading in the English markets (Rousseau and Sylla 2005, 6–9). The conclusion is inescapable: by 1825 the United States had a financial system that was gaining on the world's leaders in terms of both banking and the spread of securities markets.

Sylla et al. (2009) show how the First BUS acted decisively to rout an incipient financial crisis in 1792 when speculation in government securities and shares of the First BUS in New York led to a substantial crash and scramble for liquidity among early stock brokers. The First BUS, under the direction of Hamilton, provided the necessary liquidity at this critical moment, thereby arresting the panic. Although the crisis itself is sometimes viewed as a minor event involving only a small number of wealthy individuals, the fact remains that the BUS engaged in the type of liquidity provision that would nowadays be associated with the operations of a central bank. This suggests that the United States had at least a quasi-central bank very early in its history.

With the spread of banking came true privatization of the money creation process. The First Bank's federal charter granted it the right to issue specie-backed notes, and individual states granting bank charters also allowed this. There were no required reserve levels at the time, so these private banks could expand their issues to meet the needs of entrepreneurship, and also to maximize the profits of their owners. Since loans were often granted to bank "insiders," the probabilities of large losses tended to be small, but resources were not generally allocated through a competitive process where funds went to projects with the highest potential returns. When banks did over-issue currencies, there was always the possibility that the public might choose to redeem their notes en masse and take the individual bank down.

Interestingly, bank failures were not a concern during the time of the First BUS. One reason for this was that the First BUS began to establish branches to facilitate the collection and disbursement of the government's funds in the course of normal business. The first branches were established in Boston, New York, Baltimore, and Charleston (1792), and then later in Norfolk (1800), Washington and Savannah (1802), and New Orleans (1805) (Wettereau 1937, 278). Even President Jefferson, an original opponent of the Bank on constitutional grounds, came to appreciate the service that the Bank could provide as fiscal agent. More important for the stability of the system, however, the Bank and its branches also provided a check on paper money issues by individual state-chartered banks by collecting their notes through their ordinary operations and then deciding whether to pay them out at their own counters or to pack the notes up and return them to the counters of the issuing banks for redemption in coins. The decision was typically based on whether the First BUS could determine whether an individual bank was issuing more notes than it could redeem, and the amount of notes coming across its own counters provided a reasonable predictor. The possibility of sudden redemption by the BUS served to deter state banks from issuing too many notes. The Bank was so successful in conducting the operations that its shareholders, which included both the government and private individuals including foreigners, were paid healthy annual dividends of 7–8 percent on their stock in addition to the interest on the bonds that they had tendered for the shares.

The Bank's fortunes changed rapidly in 1811, however, when opponents in Congress stopped the financial revolution in its tracks. At that time, the Republicans, fueled by Jefferson's legacy and led by President Madison, who was even more ambivalent toward the Bank than his predecessor, caused a deadlock in the Senate on a bill to renew its charter for another twenty years just before it was to expire. The robust annual dividends were viewed by some of the Bank's opponents as evidence that a wealthy elite was unduly benefiting from use of the government's temporary balances for profit. Other opponents claimed that the Bank's federal charter as a corporation was unconstitutional in the first place. In the end, sitting Vice President George Clinton (a former Governor of New York), in his role of presiding over the Senate, cast the deciding vote against the Bank, and it ceased operations as a federal bank in 1811.

3 1812–1828

The end of the Bank could not in retrospect have come at a worse time. As British troops threatened the new republic along its Atlantic seaboard and on the Gulf coast during the War of 1812 – hostilities that included the capture and burning of Washington DC in 1814 – the federal government desperately needed funding to prosecute the campaign. Without a quasi-central bank to organize the funding efforts, the government resorted to issuing $60 million in debt directly to the public, both within the United States and abroad. It also issued $15 million in treasury notes to make up the remaining shortfall. By 1814 it had become impossible to repay these debts on schedule due to their sheer volume, and this sounded a death knell for raising the additional debt required to service existing loans.

Indeed, the United States was bankrupt in late 1814, and many banks formed in the wake of the First Bank's demise were having difficulty redeeming their own notes. It is amazing that in a time of such financial disarray General Andrew Jackson and his troops turned back the British in New Orleans in January of 1815 in the war's most decisive victory! Yet never in the history of the United States had the need for a federal bank become more apparent. Realizing the error they had made five years earlier, the Democratic-Republicans (the successors to the Democrats and now Madison's party) in Congress approved a charter in 1816 for a second federal bank, now called the Second Bank of the United States, with an even larger capital of $35 million. The new Bank began operations in 1817.

The experience of 1812–1815 reinforced the importance of having a federal bank in place to assist in financing a war. Yet the Union faced the very same deficiency again during the Civil War (1861–1865). It was exactly the need for finance at the start of the Civil War that motivated the National Banking Acts of 1863 and 1864, which set up the system of unit banks under which most of the United States operated until the founding of the Federal Reserve in 1914. But even this system, which established the Office of the Comptroller of the Currency and regulated bank note issues at the national level, represented only a partial solution to the problems of monetary control related to the absence of a central bank.

Unlike the First Bank, the Second BUS did not get the nation immediately back on track toward the modern growth that the War of 1812 had brought to a temporary halt. The Bank made many of its loans to insiders who did not invest in projects with the highest available expected returns, and shareholders continued to be rewarded with high dividends. When the Bank found its own notes coming back to its counter for redemption during a financial crisis in 1819, it mitigated these demands by sharply contracting the loan portfolio on its balance sheet. Critics claimed that the contraction, directed by Bank President and former acting Treasury Secretary William Jones, had saved the Bank at the expense of the public, a charge not lost on future U.S. President Andrew Jackson as he bided time on his Tennessee plantation near Nashville.

The appointment of Nicholas Biddle as President of the Second BUS in 1823, however, represented a sharp break with the past. Biddle took the Bank's responsibility as the nation's fiscal agent very seriously, just as Hamilton's First Bank had three decades before, but provided these services while maintaining monetary control in an economy that had grown much larger. With twenty five branch offices in operation by 1832, the Bank performed its monetary control functions once again by collecting notes of over-issuing state-chartered banks and returning them to their counters for specie. Many banks resented this form of control and charged the Bank with constricting the banking system by offering its advantages only to the few who could afford to borrow from it or hold its shares (these two groups were often one and the same). This opposition was unable, however, to gather momentum during President Monroe's second term (1821–1825) or the subsequent administration of John Quincy Adams (1825–1829). Indeed, the Supreme Court reconfirmed the constitutionality of the Bank in an 1820 ruling and then again in 1824, with the latter ruling putting the question to rest at least officially.

Yet the seeds of resistance were already being sown, as war hero Andrew Jackson won the popular vote in the 1824 presidential election but was unable to gain a majority in the Electoral College. When defeated by John Quincy Adams in a final vote taken in the House of Representatives, Jackson viewed his loss (sometimes called "The Corrupt Bargain") as unfair and typical of how the traditional elite had come to rule American politics. The response of Jackson's supporters was to form a new party – the Democrats – to address the issues of the "common" man. The Democratic party and the "populist" movement that came with it held deep suspicions of the power wrought by large banks and especially the Second BUS, and as the party swept their candidate Andrew Jackson into the White House with a landslide victory in 1828, Bank President Nicholas Biddle faced serious uncertainties about what Jackson's victory could mean for the Bank.

4 The Bank War, 1829–1834

Among the most fascinating episodes in early U.S. history, the "Bank War" waged between Jackson and Biddle from 1829–1834 is a classic example of how a political outcome can change the course of financial history. Jackson had not made public statements about his opposition to the Bank in the years leading up to his election in 1828, but Biddle suspected that Jackson's populist following might well have fostered an impression that was less than positive. Biddle tried to avert this potential problem early on by engaging the President with accounts of the great efficiency with which the Bank carried out the federal government's business. He also knew of Jackson's goal to pay down the national debt, which had accumulated from the days of Hamilton's restructuring and was exacerbated by the War of 1812, and pledged his support and cooperation to the President in seeing it through.

Biddle realized, however, that Jackson would not so easily become a supporter of the Bank during a meeting with the President at the White House during his first year in office. It was there that Biddle reports Jackson having said: "I do not dislike your bank any more than all banks. But ever since I read the history of the South Sea Bubble I have been afraid of banks." Jackson informed Biddle at the same meeting that he did not believe that the federal government held the constitutional right to charter a bank outside the "ten-mile-square" of Washington DC. At the same time, he assured Biddle that he would praise the Bank for its cooperation in paying down the debt during his Annual Address to Congress. The form of this praise, however, was not as Biddle expected and served only to raise his anxiety about the future of the Bank:

> The charter of the Bank of the United States expires in 1836, and its stock holders will most probably apply for a renewal of their privileges. In order to avoid the evils resulting from precipitancy in a measure involving such important principles and such deep pecuniary interests, I feel that I cannot,

> in justice to the parties interested, too soon present it to the deliberate consideration of the Legislature and the people. Both the constitutionality and the expediency of the law creating this bank are well questioned by a large portion of our fellow citizens, and it must be admitted by all that it has failed in the great end of establishing a uniform and sound currency.
> (First Annual Message to Congress, December 8, 1829)

After a second set of remarks near the conclusion of Jackson's 1830 address suggesting that the Bank's charter be modified to obviate constitutional objections, Biddle realized that the Bank was in trouble. By the time that Jackson's bid for re-election was underway, Biddle and his allies in Congress knew that the Bank's best hopes for renewing the charter, which was set to expire in 1836, was to raise it during the campaign. The leader of Jackson's opposition in Congress, Speaker of the House of Representatives Henry Clay, an individual with his own presidential aspirations, then hatched a plan with Biddle to push an act to renew the Bank's charter for another 20 years through both Houses of Congress in the spring of 1832. The plan was aimed at forcing the President to sign the act into law or appear foolish by using his veto power to halt legislation that "citizens" had brought forward.

The plan backfired badly on Biddle and Clay as Jackson proceeded to veto the bill on July 10, 1832 and defeat Clay in the November election with another landslide victory. The fate of the Bank was sealed at that very moment, some four years before the charter would eventually expire, and Jackson immediately embarked on a campaign to dismantle the Bank.

Many consider the "Bank War" as the battle leading up to the veto, but the events that followed are evidence that the war had only just begun. With the charter set to expire, Jackson proceeded to spend down the government's balances in the Bank and not replenish them, accomplishing what has come to be known (not quite accurately) as the "Removal of the Deposits." The new deposits were directed to the cities of New York, Boston, Philadelphia, and Baltimore, where most of the federal government's disbursements would occur. With the deposits removed, Biddle considered himself relieved from the responsibility of controlling the stock of money by collecting and redeeming notes of over-issuing banks, and this inaction, combined with inflows of specie from abroad, and particularly from Mexico, led to a severe inflation. Biddle then tried to contract the Bank's credit in response to the drain of its reserves, and this led to an apparent slow-down in business activity by 1834.

5 The panic of 1837

In the meantime, a land boom was underway. The federal government was rapidly surveying and making tracts of land available to the public in what is now considered the Midwest, as well as in states along the Mississippi and Gulf coasts. The government sold the land for a fixed price of $1.25 per acre, but speculators traveling west behind the surveyors typically purchased the most desirable tracts and resold them to "actual settlers" at significantly higher prices. Some of the speculation was fueled by the removal of the deposits, which the receiving institutions began to multiply. Enhanced flows of specie from Mexico and other points in the 1830s had also contributed to the increase in the amount of base money (Temin 1968, 268). With the federal debt fully retired (for the only time in nation's history) in 1834 and land sales rising to a fever pitch in 1834–1836, the federal government also saw its balances rising quickly, and Democrats in Congress began to clamor for something to be done about the surplus.

To these Democrats, the clear solution was to return the $34 million that had accumulated in the federal coffers to the individual states in proportion to their populations. While Jackson did not offer strong support for the plan, he acquiesced to it and signed the "Deposit Act" into law on June 23, 1836. The new law called for a "Distribution of the Surplus" to occur in four quarterly installments starting on January 1, 1837. It was a difficult reallocation to accomplish. First, it would imply a large movement of coins from the federal government's depository banks (i.e., "pet" banks) in the eastern cities to those in the interior, and this would require a rapid increase in the number of pet banks to lodge the new deposits in anticipation of their use by the states.[2] It also required the specie movements to begin in the fall of 1836 to ease the adjustment shock that would surely occur if the transfers were delayed until January 1. The end result was the assignment of 45 new "pet banks" in the second half of 1836 (Rousseau 2002), bringing the total up to 81 by December, and a movement of deposits from New York and other cities into these banks throughout the late summer and fall.

While apparently indifferent to the Deposit Act, Jackson was considerably more troubled by the land boom, and pointed to the rapid expansion of bank notes as the culprit. It is interesting that Jackson, who had been convinced by Senator Thomas Hart Benton and other advisors that hard money (i.e., coins) was the only true source of wealth and prosperity, was likely more responsible that anyone for the proliferation of notes due to his destruction of the Second Bank! Nonetheless, Jackson determined that the availability of notes was inflating land prices, and that this would end if the nation could return to a specie currency, at least when it came to land purchases. When Congress adjourned for the summer in 1836, Jackson proceeded to issue an executive order on July 11, known as the "Specie Circular," which called for the purchase of all public lands from the federal government to be made in specie starting on August 15, 1836, with exceptions for "actual settlers" continuing through December 15.

The Deposit Act and Specie Circular had several unanticipated effects that struck the U.S. economy and its banks simultaneously. First, the Specie Circular did not end the land boom as planned, or at least not immediately. Rather, potential land buyers removed specie from eastern banks and took it to the Midwest and to Alabama, Louisiana, and Mississippi to fund these purchases. Evidence from receivers of the public moneys (government representatives who collected money from those who purchased land and then placed the funds in a nearby deposit bank) saw an immediate increase in the specie component of their deposits (Rousseau 2002, 475, table 4), with more than 80 percent of deposits being made in the form of specie by the time the exception for settlers expired. This contributed to a drain of the monetary base from eastern cities and its "dislocation" in other parts of the country where the needs of merchants for gold and silver coins for international trade was much less. The Deposit Act ultimately led to $24 million in interstate specie transfers in preparation for the official Distribution of Surplus, which far exceeded the $5 million conducted specifically under the Act. The overall impact was to draw specie primarily from New York to the Southeast.

The effects on deposit banks in New York City were devastating. The Secretary of the Treasury, Levi Woodbury, frantically attempted to return specie balances from Michigan and Ohio to the northeast (Rousseau 2002, 470, table 3), but it could not be brought back quickly enough or in the necessary quantities to avert the monetary pressure that had been mounting in New York since early in the fall. By March of 1837 the deposit banks in New York City had seen their specie reserves fall from $7 million in the previous September to less than $3 million. When the reserve drain and dislocation of the base become clear, along with news of a decline in the world price of U.S. cotton, the nation's largest commercial crop, commercial bills that typically facilitated trade between the United States and its trade partners (primarily England) began to be returned for insufficient funds. These international balances needed to be settled in specie, and this was the commodity that the New York and New Orleans banks, which were at the heart of international trade, were unable to provide.

Martin Van Buren, who had been elected President the previous fall and assumed the office on March 4, 1837, chose not to repeal the Specie Circular despite a general call to do so, and when the public became aware of the precarious position in which the New York banks had found themselves, began to withdraw their deposits on May 8, 1837. What little specie was left was completely drained by May 10, and New York banks suspended payments of specie in exchange for bank notes on that day. As news of the suspension moved through the transportation network to other cities, they also suspended payments, causing a spectacular conclusion to the second-largest financial crisis in U.S. history and six subsequent years of deep recession.

The Panic of 1837 is a classic example of a Diamond and Dybvig (1983) style shift in expectations. The nation as a whole had plenty of specie to meet its obligations. The system itself was solvent. But the dislocation of the monetary base and the apparent rigidity of the political solutions determined by the Specie Circular and Deposit Act disrupted the system at its center, sending shock waves through the country and driving other banks to follow suit. If the nation had only returned their specie balances to the banks rather than hold it defensively, the crisis could have ended more quickly. Critics of the Diamond and Dybvig framework sometimes cite difficulties in pointing to the types of general shifts in expectations that can generate catastrophic demands for liquidity, but the Panic of 1837 provides a powerful case for consideration. The problems of asymmetric information that lie at the heart of Diamond and Dybvig (1983), Stiglitz and Weiss (1981), and others, and the continued relevance of asymmetric information in financial markets today, offer ample reasons for new modeling approaches to bring the partial equilibrium insights of earlier work into more sophisticated post-2008 financial modeling.

6 Free banking and the National Banking System

The recession of the 1838–1843 raised awareness of the challenges the United States would face without a central bank. With the Whig Party back in the White House in 1841, it seemed that the chartering of a new federal bank was only a matter of time. But the death of the new President and former war hero, William Henry Harrison, only a month after his inauguration and the ascension of former Democrat John Tyler brought these hopes to an end as Tyler twice vetoed the enabling legislation. The subsequent election of Democrat James K. Polk essentially brought the discussion to a close. Instead, starting with the New York Whigs in 1838, and members of both parties in the 1850s, states began putting laws into place that took the chartering of banking corporations out of the hands of politicians. These "free banking" laws implied a removal of legislative barriers to entry in favor of standardized capital requirements and lists of eligible collateral for backing note issues. Later scholars have sometimes confounded free banking with "laissez-faire" banking, which would have implied a lack a regulation, but the "free" in the term refers more directly to the destruction of entry barriers. Unfortunately, the various states did not pass uniform laws, with some prescribing lower capital requirements and a wider range of eligible securities for collateral. In Minnesota, for example, the free banking law of 1858 permitted railroad bonds to be tendered as collateral, and the decline of the underlying securities there and in other states with more lax standards led to periodic banking crises, especially in the late 1850s (see Jaremski 2010).

The outbreak of the Civil War in 1861 blunted the sting of the free banking crises, which Rolnick and Weber (1983) show did not generally end with large losses to noteholders, as preparing for war once again underscored the difficulties that the Union would face in raising funds without the assistance of a central bank and the reputational advantages it could bring to the nation's credit standing. After issuing $450,000 in fiat currency (known as the "greenbacks") that were employed in hindsight with remarkable success, the federal government passed laws in 1863 and 1864 that established the "National Banking System." The System resembled the arrangements used successfully in New York a quarter century earlier (Sylla 1969, 659), but arose more from the need to sell government debt than to improve the functioning of banks. The new laws accomplished this by specifying federal bonds as the only eligible collateral that could be tendered in exchange for bank notes, and by administering a 10 percent tax on the notes of all banks that did not join the system by October 1, 1866. Although the System ultimately failed in forcing full conversion of existing banks to national charters, it did dominate the banking system until the 1890s when a surge in new banks outside of the system ultimately surpassed national banks in number, and continued a system of unit (i.e., non-branching) banks that could not achieve the diversification necessary to consistently avoid solvency concerns (Calomiris and Haber 2014). The System also created a national pyramid of reserves in which country banks could hold a fraction of their required reserves in designated "reserve cities," which in turn held them in the "central reserve city" of New York.[3] Individual country banks could make the pyramid even more top-heavy by depositing reserves directly in New York.

With so much of the reserve base concentrated in New York and deployed as broker loans in the market for call money, the System became vulnerable to financial crises as country bankers demanded their funds at times of seasonal strain in the planting and harvest cycles. When additional unanticipated demands for liquidity arose at one of these times, reserve city banks, and especially banks in New York, would need to contract loans with sobering effects on business activity and the equity market.

The result was a currency system that had what Miron (1986) calls a "perverse elasticity," contracting when money was needed the most in the course of business and expanding when demand for money was low. This resulted in more frequent financial crises over the half century following the Civil War, though the unit nature of the system led to the establishment of clearing houses in the reserve cities that kept these crises from becoming too severe. Despite these deficiencies, Cagan (1963, 20) still comes close to the mark when he states that the United States "could not so easily have achieved its rapid industrial and commercial expansion during the second half of the nineteenth century with the fragmented currency system it had during the first half." The National

Banking System was an improvement over free banking, but did not have the capacity to tie banks together as a "system" in the way that the Federal Reserve was finally able to achieve.

Conclusion

The variegated experiments with money and financial systems that characterize the United States from its colonial days through the advent of the Federal Reserve illustrate that effective finance can be delivered in many possible forms, but that the choice of each form involves trade-offs which must be considered and dangers which must be guarded against along the way. The British North American colonies used tax anticipation notes successfully at times to provide a medium of exchange that was sorely lacking, but questions about the solvency of some colonies in the midst of excessive circulations often led to depreciations and panics. Similar concerns arose with the Continental dollars issued during the Revolutionary War as even more vague promises of redemption ended in a complete loss of confidence that turned out to be well founded. Yet the fact that colonies such as Pennsylvania and New York had successful experiences issuing such notes cannot be dismissed, and it is not obvious that it was necessary to forbid their further issuance with the Constitution (Grubb 2003).

At the same time, it is hard to argue against the idea that the "Federalist financial revolution" and the establishment of the First Bank of the United States restored the nation's credit standing and forged a financial "system" that rendered the economy less susceptible to regional economic shocks. The dissolutions of the First BUS in 1811 and Second BUS in 1836, for example, were immediately followed by periods of severe financial disarray which suggest that the benefits of central banking cannot be understated as a component of a "good" financial system (Rousseau and Sylla 2003). In this case, the establishment of a federal bank and the restoration of public credit created a demand for securities markets from which the nation never looked back.

While the Second Bank may have constricted the spread of banking and credit by the late 1820s, a situation which was corrected in the 1830s, 1840s, and 1850s after the Bank's destruction and the advent of free banking, the new systems lacked the kind of centralized control over the money creation process that the federal banks could provide (Rousseau 2015). The perpetuation of the unit banking system in the National Banking period also left individual banks and ultimately the financial system more susceptible to bad equilibria, albeit less severe ones than experienced in 1837.

The diversity of financial systems and outcomes in U.S. history highlights the challenges that economists now face in building financial models that account for the frictions that can drive the instabilities that mattered in each case. This makes developing a unified framework difficult if not impossible given our current state of understanding. But one key lesson from the United States is that the stability benefits of a strong central bank that stands ready as a lender of last resort are considerable – and that the trade-off between stability and robust growth is likely something that good policy can manage. We also know from history that the choice of regulatory intensity is critical, with excessive regulation impeding further development of the sector and lax regulation pointing towards crises. These challenges remain salient today, and innovative research in these areas is essential for improving our understanding of financial systems in a world with financial instruments and products that are increasingly complex.

Notes

[*] Professor of Economics, Vanderbilt University, Nashville, TN, USA.

[1] The notion that colonial exchange rates floated freely is not universally held. See Michener (1987) for a theoretical framework in which the colonies could have operated under fixed exchange rates with specie.

[2] Among the conditions of the Deposit Act was a stipulation that there be at least one bank assigned as a "deposit bank" in each state that chartered banks. Another limited the amount of federal deposits lodged in any individual bank to three-fourths of its authorized capital. These rules made the necessity of moving balances out of New York ahead of schedule even more apparent since some held deposits in excess of that allowed under the new law.

[3] Chicago and St. Louis joined New York as central reserve cities in 1887.

References

Cagan, Phillip. 1963. "The First Fifty Years of the National Banking System: A Historical Appraisal." In D. Carson (ed.) *Banking and Monetary Studies*. Homewood, IL: Richard D. Irwin, 15–42.

Calomiris, Charles W., and Stephen H. Haber. 2014. *Fragile by Design: The Political Origins of Banking Crises and Scarce Credit*. Princeton, NJ: Princeton University Press.

Diamond, Douglas W., and Philip H. Dybvig. 1983. "Bank Runs, Deposit Insurance, and Liquidity." *Journal of Political Economy,* 91 (3): 401–419.

Grubb, Farley. 2003. "Creating the U.S. Dollar Currency Union, 1748–1811: A Quest for Monetary Stability or a Usurpation of State Sovereignty for Personal Gain?" *American Economic Review*, 93 (5): 1778–1798.

Jaremski, Matthew. 2010. "Free Bank Failures: Risky Bonds vs. Undiversified Portfolios." *Journal of Money, Credit and Banking,* 42: 1565–1587.

Michener, Ron. 1987. "Fixed Exchange Rates and the Quantity Theory in Colonial America." *Carnegie-Rochester Series on Public Policy,* 27: 233–308.

McCusker, John J. 1978. *Money and Exchange in Europe and America, 1600–1775; A Handbook*. Chapel Hill, NC: University of North Carolina Press.

Miron, Jeffrey A. 1986. "Financial Panics, the Seasonality of the Nominal Interest Rate, and the Founding of the Fed." *American Economic Review,* 76 (1): 125–140.

Rolnick, Arthur J., and Warren E. Weber. 1983. "New Evidence on the Free Banking Era." *American Economic Review,* 73 (5): 1080–1091.

Rousseau, Peter L. 2002. "Jacksonian Monetary Policy, Specie Flows, and the Panic of 1837." *Journal of Economic History,* 62 (2): 457–488.

Rousseau, Peter L. 2006. "A Common Currency: Early U.S. Monetary Policy and the Transition to the Dollar." *Financial History Review,* 13 (1): 97–122.

Rousseau, Peter L. 2015. "Politics on the Road to the U.S. Monetary Union." In O. Humpage (ed.) *Current Federal Reserve Policy under the Lens of Economic History: Essays to Commemorate the Federal Reserve System's Centennial.* New York: Cambridge University Press, 151–173.

Rousseau, Peter L., and Richard Sylla. 2003. "Financial Systems, Economic Growth, and Globalization." In M. Bordo, A. Taylor, and J. Williamson (eds) *Globalization in Historical Perspective*. Chicago: University of Chicago Press, 373–413.

Rousseau, Peter L., and Richard Sylla. 2005. "Emerging Financial Markets and Early U.S. Growth." *Explorations in Economic History,* 42 (1): 1–26.

Sargent, Thomas J., and Bruce D. Smith. 1987. "Irrelevance of Open Market Operations in Some Economies with Government Currency Being Dominated in Rate of Return." *American Economic Review,* 77: 78–92.

Sargent, Thomas J., and Neil Wallace. 1981. "Some Unpleasant Monetarist Arithmetic." *Federal Reserve Bank of Minneapolis Quarterly Review,* 5: 1–17.

Stiglitz, Joseph E., and Andrew Weiss. 1981. "Credit Rationing in Markets with Incomplete Information." *American Economic Review,* 71 (3): 393–410.

Sylla, Richard. 1969. "Federal Policy, Banking Market Structure, and Capital Mobilization in the United States, 1863–1913." *Journal of Economic History,* 29 (4): 657–686.

Sylla, Richard. 1998. "U.S. Securities Markets and the Banking System, 1790–1840." *Federal Reserve Bank of St. Louis Review,* 80, 83–98.

Sylla, Richard, Robert E. Wright, and David J. Cowen. 2009. "Alexander Hamilton, Central Banker: Crisis Management during the U.S. Financial Panic of 1792." *Business History Review,* 83, 61–86.

Temin, Peter. 1968. "The Economic Consequences of the Bank War." *Journal of Political Economy,* 76 (2): 257–274.

Wettereau, James O. 1937. "New light on the First Bank of the United States." *The Pennsylvanian Magazine of History and Biography,* 61 (3): 263–285.

2

Monetary Transmission and Regulatory Impacts: Empirical Evidence from the Post-Crisis Banking Literature

Sanja Jakovljević, Hans Degryse and Steven Ongena

Introduction

The extent and severity of the recent financial crisis has spurred both theoretical and empirical research in the areas of macroeconomics and financial economics, largely analyzing the challenges faced or imposed by the banking sector. The focus of this chapter is placed on the empirical work that determines the relevance of banks in the monetary policy transmission mechanism, or points to regulatory and macroprudential challenges within the banking sector. While the lending channel had already been investigated prior to the crisis, the identification of the risk channel gained the attention of researchers mostly in the post-crisis literature. Empirical advances have been made on several fronts: the increased availability and use of detailed micro-data, as well as (re)development and application of several methodological approaches, have allowed solving for previously existing identification drawbacks. Furthermore, research on regulatory implications of banks' operations has proposed new ways of measuring the risk which institutions pose for the system, and suggested possible regulatory and macroprudential improvements.

Each of the chapter's sections starts with an overview of the path set by the pre-crisis theoretical and empirical research, and further describes how the post-crisis empirical literature has progressed. Although the majority of the post-crisis research deals with the causes and consequences of the recent financial crisis, the chapter also includes papers that have used the aforementioned empirical improvements to assess previous crisis episodes.

1 The lending channel

According to the traditional ("money") view of the monetary transmission mechanism, central bank open market operations that lead to reductions in banks' reserves negatively affect their reservable deposit holdings. In absence of full price adjustment, this leads to an increase in interest rates, and further to a drop in aggregate demand. The "money" view thus operates through the liability side of banks' balance sheets, and the role of banks is passive. However, monetary policy can affect aggregate demand even beyond the effect of changing interest rates. This amplifying effect works through the asset side of banks' balance sheets, due to the existence of credit frictions, and represents the "lending" view (Romer and Romer 1990). As Bernanke and Gertler (1995) explain, credit frictions give rise to the external finance premium, i.e., the difference in costs between externally-raised and internally-created funds. Main credit frictions that are the drivers of this channel are imperfect substitutability between sources of funding, both for banks and their borrowers, and asymmetric information.

One form of the lending channel – the balance sheet channel – links the external finance premium to the collateralizable net worth of the borrower: when aggregate demand is reduced (due to restrictive monetary policy), borrowers' net worth is deteriorated; hence agency costs due to asymmetric information will be higher, and so will the premium (Bernanke and Gertler 1989). Increases in the premium have a further negative effect on aggregate demand, and this amplification mechanism has been summarized by Bernanke et al. (1996) under the term "financial accelerator." Another form of this channel – the bank-lending channel – asserts that tighter monetary policy leads to loan supply reductions (Bernanke and Blinder 1992). Coupled with the inability of certain borrowers to substitute from loans to other sources of funding (Bernanke and Blinder 1988), loan supply disruptions lead to an increase in the external finance premium, and hence to a further reduction of real activity.

Early empirical evidence on the lending channel used macro-level data and relied on correlations of aggregate indicators to establish causality between money/lending and output (e.g., Bernanke 1983, Bernanke and Blinder 1988, or Kashyap et al. 1993). Other researchers warned of a post hoc fallacy embedded in such an approach (Romer and Romer 1990), or pointed to the fact that loan demand could be driving these correlations, but is unaccounted for (Bernanke and Gertler 1995, Kashyap and Stein 1995). Additionally, the effects of the lending channel might not be homogenous across firms or banks. From the point of the balance sheet channel, Kashyap et al. (1994) differentiate between bank-dependent and bank-independent borrowers; Gertler and Gilchrist (1994) and Oliner and Rudebusch (1996) consider the size of firms. For the bank-lending channel, Kashyap and Stein (1995) analyze the effects of bank size, complemented in Kashyap and Stein (2000) by liquidity considerations.

The main challenges in the empirical literature have thus been: (1) separating loan demand and supply, and (2) properly establishing causality from monetary policy to the real economy. The focus was placed on the bank-lending channel and identification of exogenous shocks that affect loan supply, but can be argued to be uncorrelated with loan demand. Examples refer to within-country studies of capital crunches (Peek and Rosengren 1995 for the US, Woo 2003 for Japan), or cross-country transmissions of capital shocks to loan supply and the real economy (Peek and Rosengren (1997, 2000) on effects of a Japanese real estate shock on Japanese banks operating in the US). Several event studies have also examined effects of bank distress, transmitted via lending relationships, on the performance of their borrowers (Slovin et al. (1993) analyze a case of bank failure in the US, Bae et al. (2002) assess exogenous shocks to banks in Korea, while Ongena et al. (2003) look at distress announcements of banks in Norway).

1.1 Identification challenges and approaches

Even though it appears that the use of bank-firm relationships can help mitigate issues with identification of supply and demand effects, Gan (2007) points to endogeneity embedded in these relationships: if there is a self-selection mechanism of firms to banks, such that unhealthy firms are paired with unhealthy banks, then adverse shocks affecting both loan supply and demand might also be running from impaired firms to impaired banks. Identification of supply and demand effects can be straightforward if loan-level data is used instead: if a firm is borrowing from two banks that are differentially affected by an exogenous shock, and receives less funds from the harder-hit bank, then causality can be argued to run from bank distress to firm performance.

The two data and methodology-related suggestions (i.e., the use of highly disaggregated data and consideration of multiple bank relationships) have later shaped much of the post-crisis empirical research. Researchers that pioneered such an approach are Gan (2007) and Khwaja and Mian (2008). They focus on exogenous shocks to loan supply in the form of drops in asset prices via real estate exposure of banks in Japan (Gan 2007), or unexpected nuclear tests in Pakistan that affected banks' liquidity due decreases in dollar-denominated deposits (Khwaja and Mian 2008). The loan-level regressions of lending growth rates use bank-level exposures to shocks and firm fixed effects to control for unobserved firm heterogeneity attributable to loan demand. As Khwaja and Mian (2008) show on the example of their exogenous shock, ignoring firm fixed effects would lead to an underestimation of the supply effects, due to a negative correlation between loan supply and demand shocks: those banks hit harder by the liquidity shock were borrowing to firms facing less difficulty in dealing with adverse shocks; hence their loan demand was less affected. The level of analysis is cross-sectional, i.e., the time dimension is handled by first-differencing pre- and post-shock data points and thus using a difference-in-difference approach.

The proposals set forward by the aforementioned authors were followed in the post-crisis research in several directions. Many analyzes were set at the bank-firm level, largely using loan-level data from credit registers; in some instances, however, the bank-firm relationships were not direct, but rather established indirectly based on firm location. Such highly disaggregated data allowed the use of the difference-in-difference approach, but demand controls were also occasionally applied in a time-varying setup. A useful feature of credit register data can be the availability of loan applications, which allow for a cleaner identification of loan demand. However, as Popov and Udell (2012) pointed out, loan application data do not account for firms that did not apply for a loan because they were discouraged; hence the entire pool of potential borrowers is not considered. For that reason, firm-level surveys containing indicators on the discouragement to apply for a loan provide a better possibility to identify loan demand factors. Similarly, bank-level surveys on lending standards of banks contain indicators of loan demand that, as argued by Ciccarelli et al. (2014), refer to the entire specter of potential borrowers. They have also been applied in vector autoregressive models (VAR) to point to macroeconomic implications of loan supply and demand shocks. Finally, disequilibrium models, which rely on the definition of credit-constrained firms and estimate a system of loan demand and supply functions, have also been used to separate loan demand and supply factors.

The main focus of the post-crisis empirical literature has been placed on identification of the bank-lending channel, although some papers also consider the separation of the bank-lending and the balance-sheet channels, mostly using country-level data (e.g., Kalemli-Ozcan et al. 2015, Ciccarelli et al. 2014). Part of the literature focused on one of the underlying assumptions of the bank-lending channel: substitutability between loans and corporate papers/bonds for firms. The relevance of this friction, as Carvalho et al. (2013) demonstrate using data on US syndicated loans, is that bank dependence can aggravate spillovers of bank distress to the real performance of firms. While Kashyap et al. (1993) already found evidence of substitutability using aggregate data, similar results have been documented by Adrian et al. (2013) or Becker and Ivashina (2014) with US firm-level data. The latter authors additionally argue that this substitutability is indicative of a contraction in loan supply: firms that substitute from loans have a positive demand for loans (conditional on issuing new debt); hence, if they switch to bonds, loan supply must have reduced. Identification of loan supply was also enabled through exploration of exogenous shocks and "natural experiments," and the recent crisis offered ample opportunities to explore shocks that can be argued to be unrelated to loan demand: the dry-up of the wholesale funding market, issues at the interbank markets, and the rise of the sovereign debt crisis. Moreover, exogenous shocks to loan demand have also been identified.

1.2 Bank-lending channel: effects on lending

In this part of the literature, the empirical strategy was strongly driven by Gan (2007) and Khwaja and Mian (2008), i.e., controls for (un)observed firm-level heterogeneity were used largely in a difference-in-difference approach, occasionally in a time-varying manner. The predominant source of data were credit registers, combined with bank- and firm-level information. What tends to be a striking difference in these studies is the representativeness of the credit register data, i.e., the variation in the reporting threshold.

In an analysis of the transmission mechanism strength, Jiménez et al. (2012) use monthly loan application data from the Spanish credit register (with a low threshold of reporting of 6,000 EUR), and employ time-varying controls for (un)observed firm heterogeneity, i.e., firm-month fixed effects. Their results suggest that fewer loan applications were granted in times of higher short-term policy rates or low GDP growth, and this effect is stronger for banks with low capitalization or liquidity levels.

Several papers analyze the transmission of supply-related shocks to lending. Iyer et al. (2014) exploit quarterly loan-level data from the Portuguese credit register, with an even lower reporting threshold of only 50 EUR. The authors find that higher pre-crisis interbank exposure led to larger drops in growth rates of corporate loans, as well as less creation of new lending relationships and more terminations of existing ones. The effects appear stronger for smaller firms, which could not substitute easily for other sources of credit, and for banks with more non-performing loans. An identical methodological approach was used by Bonaccorsi di Patti and Sette (2012), who focus on monthly credit register data for Italy (with a reporting threshold of 75,000 EUR) to assess how banks' access to the interbank market and securitization practices influence the volume and price of loans, as well as the approval of new loans. These factors mattered for pre-crisis lending, while in 2008 banks' capitalization levels influenced their sensitivity to funding shocks. Again on the Italian sample of loan-level data and also using firm-month fixed effects, Bofondi et al. (2014) study the differential impact of the Italian sovereign debt crisis on lending practices of domestic and foreign banks, with the effect presumably weaker for foreign banks. Results confirm that domestic banks reduced their supply of credit more than foreign banks, both at the intensive and extensive margin. Contrary to previous analyzes, Berg and Schrader (2012) identify events that affect loan demand: volcanic eruptions in Ecuador. The authors use loan-level data on loan applications and approvals from a microfinance institution, and match aggregate shocks to firms based on their location. Their findings suggest that in periods of increased loan demand (due to a recent eruption episode) the probability of having a loan application approved is reduced, which is indicative of loan supply contractions.

Exogenous shocks to banking can also be transmitted from other countries. Puri et al. (2011) differentiate between German savings banks based on their ownership stakes in those German regional banks (Landesbanken) with direct exposure to the US subprime crisis. Using quarterly data on loan applications and their granting outcome, they find that affected banks rejected more applications compared to non-affected ones, although both types of banks faced a similar decline in loan demand, i.e., less applications filed (see also Ongena, Tümer-Alkan et al. 2015). Schnabl (2012) exploits an exogenous shock of the Russian default in 1998 to Peruvian banks, using loan-level data on corporate loans with a reporting threshold of 5,000 USD. The analysis is performed for loans at the bank-bank level (with bank-loan-originator and bank-loan-recipient effects to control for the average changes in loan supply and demand) and at the bank-firm level (with firm fixed effects to control for loan demand). The liquidity shock led to a reduction of both international interbank lending, and lending from Peruvian banks to domestic firms. Cetorelli and Goldberg (2011) investigate how loan supply of banks in Europe, Asia and Latin America was affected by liquidity shocks originating in developed countries, which were transmitted via reduced lending of foreign banks (both directly and via affiliates), and of domestic banks due to reduced interbank lending. Using a sample of country-level bilateral lending data, the authors validate all three predicted channels of liquidity shock transmission. De Haas and Van Horen (2013) analyze how cross-border lending was affected by the onset of the financial crisis, i.e., the Lehman collapse. They use a sample of syndicated loans at the bank-country and bank-firm level to investigate whether there was a sudden stop in cross-border lending from a bank to its cross-border borrower, and how the volume and number of loans from a bank to another country were affected. Proximity mattered for both the degree of lending and the decision of banks to withdraw from specific markets.

1.3 Bank-lending channel: effects on real outcomes

In this line of research, the focus was placed on investment outcomes of firms; other research looks at export or employment effects. The use of bank-firm level data, augmented with real activity indicators for firms, was the main advancement of the post-crisis research. The difference-in-difference approach dominated in the analyzes.

1.3.1 Investment

Amiti and Weinstein (2013) use loan-level data on Japanese listed firms to assess the impact of bank supply shocks on aggregate loan supply and investment. They point to an inefficiency of the methodology employed by Khwaja and

Mian (2008): a regression of loan growth rates on bank-and firm fixed effects ignores adding-up constraints, i.e., the fact that firms only borrow more if one of the banks lends more and vice versa. For this reason, the predicted and actual growth rates of lending are not highly correlated. Therefore, the authors also include shares of a firm in a bank's lending portfolio, and vice versa, in the estimation procedure of the aforementioned regression. The estimated bank shocks are then aggregated at firm-level and used with estimated firm shocks to assess their effect on the investment/capital ratio of firms. The bank-and firm shocks are further aggregated and used together with industry and common shocks to assess their influence on aggregate investment and lending. Results suggest that bank supply shocks have substantial effects on firm and aggregate-level investment.

Chava and Purnanandam (2011) exploit the same shock as Schnabl (2012) to the capital positions of US banks and assess the subsequent effects on their corporate borrowers. Their analyzes at the loan and firm level show that affected banks reduced loan volumes and increased lending rates after the crisis, while firms borrowing from more affected banks decreased their capital expenditures and had more pronounced reductions in profitability. Cingano et al. (2013) focus on the dry-up of the interbank market liquidity as a shock to the loan supply in Italy, and use Italian credit register data to assess how investment choices of firms were affected. The authors use firm fixed effects as controls for unobserved heterogeneity; however, as Jiménez et al. (2011) pointed out, within-firm analyzes do not account for the possibility of firms to switch banks when experiencing credit drops. For that reason, the firm-level regressions of credit growth on interbank exposure use unbiased estimates of firm fixed effects from the equivalent loan-level regression. They find that firm investment has been affected by the exposure of their banks to the interbank market shock. Balduzzi et al. (2014) show that Italian banks were also hit by the financial and sovereign crisis through increases in their CDS spreads and lowered equity valuations, which had an adverse effect on borrowing firms in terms of their investment, employment and bank debt. A vast majority of firms in their sample borrow from just one bank; hence, the identification strategy includes both firms borrowing from multiple banks and a sample of single-bank firms.

The issue of plentiful single-bank firms was also present in the analysis by Ongena, Peydró et al. (2015), who investigate the heterogeneity of international shock transmission from the financial to the real sector in a cross-country study. The authors use bank-firm level data from 14 countries in Eastern Europe and Central Asia. Their analysis at bank and firm level shows that lending contractions were larger among domestic banks borrowing internationally and foreign-owned banks, and their borrowers experienced greater contractions of their real and financial performance. The authors also emphasize that combining loan-level data with firm characteristics is an important part of identification, since solely firm fixed effects, when used as controls for loan demand shocks, do not contribute much to the variation of lending once firm characteristics are accounted for.

1.3.2 Export

Zia (2008) uses matched bank-firm data for a sample of Pakistani textile exporters and investigates how the abolishment of the credit subsidy program for yarn producers affected export of firms; privately owned firms were affected more compared to public ones. Del Prete and Federico (2014) use loan-level data from the Italian credit register on a sample of Italian exporters, which further contains information on the purpose of the loan: import, export or ordinary loans. The difference-in-difference approach using time-varying demand controls suggests that loan contractions were larger for banks hit harder by foreign funding shocks, and more so for ordinary loans. Amiti and Weinstein (2011) match loan-level with export data for Japanese listed firms based on the main bank providing trade credit, and find that financial health of banks affects export performance of firms. Since exporters might be borrowing from healthy or unhealthy banks that could also have a preference for a specific industry, the use of industry-time fixed effects takes out supply and demand-driven shocks common to exporters, so within one industry the export performance of exporters borrowing from an unhealthy bank can be identified. Paravisini et al. (2015) use loan-level data from the Peruvian credit register to assess the impact of credit constraints on the volume of export and entry/exit decisions to export on a firm-product-destination level. The authors estimate how lending from banks that were differentially affected by the capital flow reversal affects changes of export of the same firm, accounting also for non-credit shocks from the exporting markets with product-destination fixed effects. While changes in credit conditions affected the intensive margin, they didn't influence the extensive margin.

1.3.3 Employment

Bentolila et al. (2013) use data from the Spanish credit register and split firms into groups of those that borrowed relatively more from healthier banks (i.e., those that did not require a capital injection during the crisis) and those that borrowed more from weak banks (i.e., those that were bailed out). These groups do not overlap due to the condition that firms borrowing from weaker banks could not switch to healthier banks. Based on a difference-in-difference and matching approach, they find that employment reductions were additionally lower in firms linked to weaker banks. Chodorow-Reich (2014) looks at US firms that obtained syndicated loans and analyzes the effect of lender health, defined post-crisis, on lending and employment outcomes of firms that had pre-crisis relationships with healthy versus unhealthy lenders. As in Bentolila et al. (2013), links to unhealthy banks have detrimental influence on employment, but the result also depends on the size of the firm: while this negative impact is pronounced among SMEs, there appears to be no effect on larger firms.

1.4 Alternative data sources and modeling strategies

1.4.1 Firm and bank surveys

As previously mentioned, firm-level surveys were used to directly control for firms that do not apply for loans due to discouragement; hence they would not be accounted for either in credit register or loan application datasets, although they are also affected by loan supply shocks. Popov and Udell (2012) use survey data from the Business Environment and Enterprise Performance Survey (BEEPS) for 16 countries of Central and Eastern Europe, in order to assess how credit supply was affected by financial health of banks. The authors separate demand from supply by considering firms that report being credit-constrained due to supply-driven factors (i.e., their loan application was rejected, or due to loan specifics such as the interest rates, collateral requirements etc., which acted discouraging on the firms). Bank-firm relationships were determined based on the presence and market share of banks in localities of the respondent firms. The authors find that riskier firms and firms with fewer tangible assets are more hardly hit by capital shocks to banks, indicative of "flight to quality" effects. A similar approach was applied by Presbitero et al. (2014), who use survey data on loan applications and approvals for Italian manufacturing firms, in order to identify factors leading to bank withdrawals from local markets. Bank-firm links were created combining data on bank branch openings and closures with firm-level data based on firm location; credit-rationed firms are defined as those which did not obtain the desired amount of credit. The authors find that credit reductions were larger with more functional distance within a bank. Contrary to findings by Popov and Udell (2012), adverse effects were stronger for financially healthier firms, suggesting a "home bias" rather than a "flight to quality" effect. It is difficult to assess whether these differing results stem from varying definitions of rationing, or are due to imprecisions in establishing bank-firm relationships.

Surveys on bank-lending standards have already been used in the pre-crisis research that found a significant effect of these standards on aggregate lending and output, using VAR models (see Lown et al. 2000, Lown and Morgan 2002, 2006). These surveys regained attention in post-crisis analyzes, since changes in lending standards can be attributed to bank-driven factors and are therefore possible indicators of loan supply changes. Surveys also contain information on loan demand estimates by banks, which have been used as loan demand controls, and on reasons for changes in both lending standards and loan demand. In a country-level setting for Euro area countries, Hempell and Kok Sørensen (2010) use quarterly Bank Lending Survey (BLS) data and show that supply constraints related to bank liquidity and access to market finance had an especially pronounced effect on loans supplied during the crisis. Among within-country studies, van der Veer and Hoeberichts (2013) use BLS data for eight Dutch banks, and find that tightening of lending standards permanently reduces lending. Del Giovane et al. (2011) combine loan-level data on lending of Italian banks to the non-financial firms with responses of banks to the BLS survey. While both demand and supply factors contributed to the movements of the growth rates of lending, the effect of supply was mostly pronounced after the Lehman collapse. When the BLS responses on changes in supply lending standards are interacted with causes of their changes, it can be seen that the majority of the changes in growth of lending can be attributed to costs related to the capital position of banks, and the rest to perceptions of risk. For the case of United States, Demiroglu et al. (2012) use the quarterly Senior Loan Officer Opinion Survey (SLOOS) to analyze whether lending standards of banks have a differential impact on lending to private and public firms. Differentiation between loan supply and demand is based on the substitutability between bank loans and trade credit in times of credit tightening, which is argued as indicative of a supply effect. They find that periods of tightened lending standards are associated with reduced lending to private firms at the extensive margin, but there appears to be no difference between firm types at the intensive margin.

1.4.2 VAR

Closely related to the use of data on bank-lending standards in post-crisis research was the application of the VAR methodology, in order to disentangle demand from supply effects and assess their interactions with macroeconomic outcomes. Ciccarelli et al. (2014) apply a VAR setup to measure the influence of the monetary policy rate on real activity via bank-lending and balance-sheet channels, using country-level data from bank lending surveys in Europe (BLS) and the US (SLOOS). The advantage of these surveys is that their design allows separating between changes of loan demand, net worth of borrowers (balance-sheet channel) and net worth of lenders (bank-lending channel). The authors find evidence of the lending channel in monetary policy transmission to GDP and inflation: while the bank-lending channel is not significant in the US context, it is relevant in Europe, along with the demand channel. Using the SLOOS survey on US banks, Bassett et al. (2014) construct an indicator of loan supply based on bank-level responses on lending standards, and incorporate it into a VAR model to assess its broader economic impact. Their results suggest that negative loan supply shocks lead to reductions in the borrowing capacity of the non-financial sector, real GDP reductions, widening of credit spreads for the corporate sector and monetary policy easing.

Identification of loan supply shocks and their macroeconomic effects can also be assessed with VAR models using sign restrictions. As Hristov et al. (2012) demonstrate, sign restrictions are used in the following way: if drops of loan volumes are observed, they can be attributed to a loan supply shock if there is a simultaneous increase in the loan rate; conversely, simultaneous drops in loan rates are indicative of loan demand shocks. Their cross-country results point to a large influence of loan supply shocks on lending and output growth during the recent crisis, but also emphasize the substantial heterogeneity of this effect across countries.

1.4.3 Disequilibrium models

Several papers have also used disequilibrium models to separate loan demand from loan supply effects on observed loan volumes, and the crucial identification point of such models is separating credit-constrained firms from those that do not face these constraints. Examples of such models can also be found in the pre-crisis literature (e.g., Ogawa and Suzuki 2000 or Atanasova and Wilson 2004), and their resurgence can be attributed to challenges with the lending channel identification in times of the recent crisis.

Disequilibrium models are specified using a system of a loan demand equation, loan supply equation and a transaction equation: in the first two equations, determinants of loan demand and supply include firm characteristics and the interest rate of the loan; exclusion restrictions are also imposed that allow for identification. The transaction equation imposes that the observed loan volumes are the result of the minimization of the loan demand and supply functions. Firms for which loan supply exceeds loan demand are considered as unconstrained, while the opposite holds for constrained firms. Carbó-Valverde et al. (2013) apply the model to a sample of Spanish SMEs using firm-level data. They find that credit-constrained firms depended relatively more on trade credit than bank loans during the crisis, while less credit-constrained firms were dependent on bank loans. Results point to the existence of a degree of substitutability between bank debt and other types of external finance for capital expenditures. From a similar sample of firms, Carbó-Valverde et al. (2012) find that firms faced less credit constraints prior to the crisis if they had a relationship with a bank involved in securitization activities, while in crisis periods the type of securitization had a differential effect on credit rationing. Kremp and Sevestre (2013) apply the model to identify whether reductions in lending to French SMEs were due to decreases on the side of demand or supply. As opposed to other papers' estimation procedure, they also take into account the firms without loans, i.e., firms for which the interest rate is unobservable and hence cannot be included in the loan demand and supply equations. The authors find that demand factors were decisive for observing lower loan amounts to SMEs.

Using responses of Italian banks to the BLS survey, Del Giovane et al. (2013) extend a standard three-equation disequilibrium model by adding a price adjustment equation and two equations that assume the following: (1) banks are reporting tightening of lending standards in times of excess loan demand in the market, (2) banks report easing of lending standards in times of excess loan supply in the market. This system is further reduced to two equations for loan demand and supply. Differentiating between periods of the global crisis and the sovereign debt crisis, the authors find that weak loan demand contributed equally in both periods to movements of loan volumes and bank mark-ups, but the effects of tightening standards were more pronounced in the sovereign debt crisis period, largely due to conditions in banks' balance sheets and funding.

2 The risk channel

Another channel of monetary policy transmission that is credit-related, but links more closely to the attitude of banks towards risk is the risk-taking channel, the term first introduced by Borio and Zhu (2012). In the pre-crisis period it was analyzed from a theoretical perspective, but received substantial empirical validation in the post-crisis era of low interest rates.

Rajan (2005) warns of the possibility of a "search for yield" by institutions with more long-term liabilities (hedge and pension funds) in times of low interest rates, which can also feed into asset price increases and risky investments. Another channel through which low policy rates feed into risk-taking behavior of banks is through the effect on leverage ratios of banks. Adrian and Shin (2009) point to the procyclicality of banks' leverage. In times of low policy rates, as De Nicolò et al. (2010) explain, increases in asset prices will lead to an increased demand for assets, which feeds into an additional increase of asset prices. They also point to a third possible channel: asset substitution that leads to a reduction of the share of safer assets in banks' portfolios and an increase in demand for risky assets. Capitalization levels may influence which of the channels dominates: overall, banks with lower capital levels will tend to take less risk than better-capitalized banks. Motivated by increases in asset prices and credit at the onset of the recent crisis, Acharya and Naqvi (2012) develop a model where, in abundance of liquidity, loan officers have motivation to originate more loans, giving rise to increased risk-taking by banks and the creation of asset bubbles. Modeling the effects of low interest rates on incentives of banks to take risks, Dell'Ariccia et al. (2014) consider changes in banks' leverage and monitoring efforts when their capital structure is either fixed or endogenously determined.

In empirical assessments of the risk-taking channel, the data largely originate from bank lending surveys or credit registers. Maddaloni and Peydró (2011) use quarterly data from the US and 12 European countries' lending surveys on changes in lending standards, and find that low short-term and long-term rates led to loosening of lending standards. For the US, Buch et al. (2014) use survey questions on risk assessments of new loans from the Survey of Terms of Business Lending (STBL). The authors use the factor-augmented vector autoregressive model (FAVAR) that allows including rich data on bank characteristics and can thus account for two-way feedbacks between macroeconomic, monetary and banking variables. The authors find heterogeneous responses of banks to low policy rates: while for small banks there is evidence of a risk-taking channel, it is not so for large and foreign banks. However, both small and foreign banks respond to prolonged low rates by more risk-taking. De Nicolò et al. (2010) augment their model specification by empirical validations using quarterly data from the STBL survey, and Call Reports of US banks. Overall, low policy rates induce more risk-taking by banks. Dell'Ariccia et al. (2013) use the STBL survey to find that the ex-ante risk-taking (measured by the internal rating of banks' portfolios) is negatively correlated with policy rates, but the effect is less pronounced for banks with lower capital levels, or during periods of capital erosion (i.e., in crises).

Another set of country-level studies uses credit register data, augmented with bank- and firm-level information.[1] Gaggl and Valderrama (2010) focus on the period when ECB refinancing rates were "too-low-for-too-long," and analyze how monetary policy changes affected the risk positions of corporate borrowers from Austrian banks. The lower limit for credit reporting is 350,000 EUR, while the used sample of firms tends to be skewed towards larger and sounder corporate borrowers. The authors compare the ECB refinancing rate to the Austrian Taylor rule, and find that the average expected default rate of the borrowing portfolio of Austrian banks is higher in periods that are expansionary according to the Austrian Taylor rule. Jiménez et al. (2014a) use Spanish credit register data on loan applications and approvals to address more directly how the composition of loan supply is affected by low monetary policy rates. Since these changes occur at the bank-firm level, for clear identification it is necessary to separate those changes from changes in volume of supply (i.e., at the bank level) and from changes in quality and volume of demand (i.e., at the firm level). This is achieved using bank-time and firm-time fixed effects that control for (un)observed heterogeneity. The authors find evidence of the risk-taking channel that also varies according to capital levels of banks. Using credit register data for Bolivia, Ioannidou et al. (2014) analyze the additional effect of low interest rates (exogenously transmitted from the US) on loan pricing decisions of banks. Their results using newly approved loans suggest that reductions of overnight rates make it more likely for banks to grant loans to ex-ante riskier firms, as well as to firms that are likely to default. Also, both the expected returns and loan price per unit of risk decline as the overnight rate drops, suggesting that this effect is supply-driven, even more so for banks facing moral hazard issues.

A few studies have attempted to draw cross-country conclusions. Altunbas et al. (2014) look at listed banks in 14 EU countries and the US using data from banks' balance sheets and an indicator of a bank's probability of default, and find evidence of a negative effect of prolonged low interest rates on bank risk (i.e., the risk increases). Delis and Kouretas (2011) focus on annual data of banks from 16 Euro area countries and show that long periods of low interest rates led banks to take more risky positions. However, it appears that banks with higher levels of capital exhibit a stronger relationship between low policy rates and risk-taking, as do banks with more off-balance sheet items.

3 Financial innovation: securitization issues

The previous section emphasized how the risk-taking behavior of banks can adversely affect the broader economic setting. A specific financial innovation that closely relates to how banks tackle risk is securitization. Although this practice can have positive effects on banks' lending capacities, it may also reduce the ability of monetary authorities to affect banks' loan policies, as documented by Loutskina (2011).

A renewed methodological approach that was used to assess whether securitization could have also posed a threat to the stability of the system is regression discontinuity design. Keys et al. (2010) apply it in order to assess how higher availability of securitization by lenders has influenced the default probabilities of their borrowers' portfolio, looking at the US subprime mortgage market. The authors apply a broadly used measure of the credit quality of borrowers in the US (FICO score) and an *ad hoc* cut-off level of this measure stemming from underwriting guidelines: lenders should not lend to borrowers with a FICO score below 620. Results indicate that the threshold value indeed represented a discontinuity point, since the number of securitized loans is higher slightly above the threshold. Further analyzes suggest that, among securitized loans, those with a score slightly above the threshold also have a higher probability of default than those loans with scores below the threshold.

However, Bubb and Kaufman (2014) criticize the main assumption of Keys et al. (2010) that the incentive for lenders to differentially screen borrowers around the FICO threshold is driven purely by securitization reasons. The authors instead argue that the guidelines for the FICO threshold apply to the decision to originate a loan, not to securitize it. They extend their dataset to non-securitized loans as well, and find that the number of originated loans and their default probabilities indeed vary around the credit score threshold, but the securitization rates are not affected. For that reason, they warn that policy implications of previous papers, which criticized securitization practices due to their potential moral hazard issues, were highly inappropriate.

4 Regulation, macroprudential policy and their challenges

Beside addressing possible stability concerns related to financial innovations, some of the main challenges of regulatory and macroprudential policy are to tackle the issues of procyclicality and risk of the system, while at the same time assuring coordination of regulatory policies in a cross-country context. These concerns gained attention in the post-crisis research, and the empirical literature in this area has also been increasingly using disaggregated data.

4.1 Procyclicality

The issue of procyclicality of capital buffers (i.e., the negative co-movement of capital buffers and the business cycle) has been identified even prior to the crisis, and further evidence continues to be provided.[2] This procyclicality is due to a negative relationship between capital buffers held at banks and the business cycle: since banks could reduce their capital holdings in times of economic upturn, this could lead to declines of their lending activity and to a reduction of economic growth. However, researchers have also emphasized the potential procyclicality of capital-based regulatory measures, which could exacerbate the inherent procyclicality in the banking sector. Since capital requirements are positively related to the risk in the economy, and risk is higher in times of recessions, higher capital holdings by banks can result in further drops of economic activity. Such effects were already identified for Basel I capital regulations (e.g., Jackson et al. (1999) for developed countries, Chiuri et al. (2002) for developing countries), and for Basel II risk-sensitive capital requirements (e.g., Fabi et al. 2005, or Jokipii and Milne 2008, see also VanHoose 2008 for a survey). Recent papers assessed the impact of the standardized (SA) and internal ratings-based (IRB) approach for credit risk assessment on lending activity of banks, using bank-firm level data. Behn et al. (2014) analyze the introduction of the IRB system in Germany in 2007, and use credit register data (with a reporting threshold of 1.5 million EUR) to show that banks implementing the IRB approach, characterized by lower capital charges, increased their lending more compared to banks using the SA approach. Fraisse et al. (2015) investigate the impact of the switch from Basel I to Basel II capital requirements on lending by banks, using loan-level data from the French credit register (with 25,000 EUR as the reporting threshold), and find evidence of loan size increases resulting from a decrease in capital requirements. In addition to using loan-level data, both papers also employ a difference-in-difference approach proposed by Gan (2007) and Khwaja and Mian (2008).

Regulatory changes suggested within the Basel III framework aim at reducing the procyclical effect of capital, introducing the countercyclical capital buffer as a relevant macroprudential tool. Several analyzes assess the potential efficiency of this instrument. While Tabak et al. (2011) and Shim (2013) support its introduction, Francis and Osborne (2012) use a simulation exercise to show that the effectiveness of this policy measure will depend on how successfully banks can be averted from fulfilling capital requirements via lower-quality capital. Grosse and Schumann (2014) point out that countercyclical measures will be effective only if there are no underlying internal reasons that could drive the negative relationship between capital buffers and the business cycle, such as risk aversion or rating schemes. Basten and Koch (2015) provide an analysis of the first case of countercyclical capital buffer implementation, which took place in Switzerland in 2013. The authors use detailed data from a Swiss mortgage broker that allows them to separate between mortgage demand and supply, and therefore assess the capital requirement shock on mortgage supply and mortgage rates charged. Results suggest that risk-weighting schemes did not affect the lending activity towards very risky borrowers in light of the countercyclical capital buffer implementation; hence, the increased capital requirements did not discourage risky lending. Jiménez et al. (2014b) analyze dynamic loan loss provisioning, a countercyclical forward-looking policy instrument that has been in place in Spain even prior to the Basel III policy suggestion, and show its effectiveness in reducing lending cycle fluctuations. The authors use three changes in the dynamic provisioning design (its introduction and two instances of adjustments) as shocks to bank capital, and assess its effect on lending using detailed loan- and loan-application-level data, using again a difference-in-difference approach.

4.2 Risk

Another relevant point for macroprudential policy is improving identification of (systemic) risk and prevention of excessive risk-taking by financial institutions. Post-crisis research oriented towards measuring systemic risk has largely focused on assessing contributions of individual institutions to the overall risk of the system. Adrian and Brunnermeier (2014) offer an extension of the traditional measure of individual risk of an institution (value at risk) and relate it to the entire sector by considering the conditional/contagion/co-movement value at risk: CoVaR. This measure of risk for an institution, relative to the system, is the value at risk (VaR) of the system conditional on the institution being under distress. The difference between this measure and CoVaR conditional on the normal state of the institution is the individual marginal contribution of an institution (in a statistical rather than causal way) to the overall systemic risk, or ΔCoVaR. Based on another standard measure of firm risk – i.e., on the expected shortfall – Acharya et al. (2012) define the systemic expected shortfall (SES) measure of the contribution of a financial institution to systemic risk as the propensity of an institution to be undercapitalized in circumstances when the whole banking system is undercapitalized. This measure can then be related to the marginal expected shortfall (MES) of an institution, i.e., its losses that are in the tail of the sector's loss distribution. Based on the latter measure in its long run form, as well as bank size and degree of leverage, Brownlees and Engle (2015) introduce the SRISK index, i.e., the expected undercapitalization of a bank conditional on a prolonged market decline. Tarashev et al. (2010) use a game theory concept of the Shapley value to allocate system-wide risks to individual institutions, arguing that such a methodology is equivalent to assessing contributions of individual institutions to systemic risk.

Looking at post-crisis research that simultaneously uses several systemic risk measures, Brunnermeier et al. (2012) use both the ΔCoVaR and SES measures for US publicly traded bank holding companies and show, using the difference-in-difference methodology, that contributions of individual institutions to systemic risk following the Lehman collapse were higher for those institutions where non-interest income mattered more. In a cross-country setting, Laeven et al. (2014) use the ΔCoVaR and SRISK measures of systemic risk, as well as standalone risk of banks as proxied by market returns, to assess how much variation of those measures can be attributed to specific bank characteristics. Using data on publicly traded financial institutions from 56 countries, and controlling for country-specific factors, they find that large banks and, to some extent, banks with lower capitalization levels, impose more systemic risk.[3] Both papers point to the finding that correlations between different measures of systemic risk can be low, since they incorporate various features into the assessment of systemic risk, but also that common consideration of the measures can be informative.

4.3 Regulatory spillovers

Another point relevant for efficient regulatory policies is to ensure cross-country coordination of implemented practices. Examples of regulatory spillovers in the empirical literature show that differences in regulatory restrictions can give rise to risk-shifting incentives of banks operating internationally. Aiyar et al. (2014a, 2014b) analyze differing responses of regulated domestic banks and non-regulated foreign branches, in terms of loan volume, to bank-specific capital requirement changes in the UK. The authors control for demand changes by including measures of both sectoral and bank-specific loan demand, and find that foreign branches partially offset the effect of increased capital requirements on reductions in loans supplied, due to credit substitution between regulated and non-regulated business units of foreign banks. Fidrmuc and Hainz (2013) focus on a specific example of differences in regulations regarding reporting standards on borrower quality in Germany and Austria, with stricter reporting standards in the former country. A difference-in-difference methodology indicates that, during the period when regulatory differences existed, cross-border lending by Austrian banks increased, and German firms close to the Austrian border had a higher overall probability of obtaining a loan from banks in either of the two countries. It could also be the case, however, that increased lending in times of looser regulatory policies may be related to lower lending standards and more risk-taking by banks. Such effects were found by Ongena et al. (2013), who analyze the spillover effects from multinational banks to their host markets, when faced with stricter regulation in their home country relative to the host market. Similarly to Popov and Udell (2012), firm-survey data on loan applications and approvals was matched with banks based on locality. The estimation procedure also includes controls for potential self-selection of foreign banks into specific host countries and local markets. Results suggest that risk-taking practices abroad by banks facing regulations that allow more competition or restrict activities in the home market are even more pronounced when regulation at home is inefficient.

Conclusion

The main challenges in the empirical literature on the monetary transmission mechanism have been to separate loan demand and supply, and to establish causality from monetary policy and regulation to the real economy. The post-crisis literature has been able to employ improved identification methods due to the availability of new detailed datasets and the occurrence of exogenous shocks. The conclusions reached also point to various dimensions affected by the mechanism, from lending outcomes to real effects. The recent financial crisis has led to an emergence of several shocks that affected loan supply by banks, which made the academic discussion on the bank-lending channel more comprehensive compared to the pre-crisis research. As a consequence of the policy response to the crisis, we now face a long period of extraordinary low interest rates, which seems as a favorable environment for the risk-taking channel to be gaining more importance.

Although many of the identification issues that characterized previous research have undoubtedly been removed in the post-crisis literature, several caveats that concern data features still remain. For instance, threshold values for reporting to credit registers in some countries might be overly restrictive. Lowering these reporting requirements would help to identify impacts for the smaller firms, especially since they have fewer loan substitution possibilities, and may only be able to move to informal financing sources. More precise variable definitions could also be specified, e.g., when measuring loan availability or establishing bank-borrower relationships. Without such data alignments, international comparisons of obtained results are hard to make, and this is even more true if regulatory practices and their assessments are further considered. Nevertheless, policy-making has benefited immensely from the recent empirical contributions, while further progress can still be made in aligning the macro- and financial aspects of banking and addressing the heterogeneity of banking systems.

Notes

* Sanja Jakovljević and Hans Degryse acknowledge financial support from FWO-Flanders under contract G.0719.13.

1. Also see further references in Ioannidou et al. (2014) and Jiménez et al. (2014a).
2. See Jakovljević et al. (2015) for a review.
3. See also Mutu and Ongena (2015) on the impact of policy interventions on systemic risk across banks.

References

Acharya, V., Engle, R., and Richardson, M. 2012. "Capital Shortfall: A New Approach to Ranking and Regulating Systemic Risks." *American Economic Review Papers and Proceedings*, 102 (3): 59–64.

Acharya, V., and Naqvi, H. 2012. "The Seeds of a Crisis: A Theory of Bank Liquidity and Risk-Taking Over the Business Cycle." *Journal of Financial Economics*, 106 (2): 349–366.

Adrian, T., and Brunnermeier, M.K. 2014. *CoVaR*. Federal Reserve Bank of New York, New York NY.

Adrian, T., Colla, P., and Shin, H.S. 2013. "Which Financial Frictions? Parsing the Evidence from the Financial Crisis of 2007 to 2009." In *NBER Macroeconomics Annual 2012*, Acemoglu D, Parker J & Woodford M (eds). University of Chicago Press, Princeton NJ, 159–214.

Adrian, T., and Shin, H.S. 2009. "Money, Liquidity and Monetary Policy." *American Economic Review Papers and Proceedings*, 99 (2): 600–605.

Aiyar, S., Calomiris, C.W., and Wieladek, T. 2014a. *How Does Credit Supply Respond to Monetary Policy and Bank Minimum Capital Requirements?* Bank of England, London.

Aiyar, S., Calomiris, C.W., and Wieladek, T. 2014b. "Does Macro-Prudential Regulation Leak? Evidence from a UK Policy Experiment." *Journal of Money, Credit and Banking*, 46 (1): 181–214.

Altunbas, Y., Gambacorta, L., and Marquez-Ibanez, D. 2014. "Does Monetary Policy Affect Bank Risk?" *International Journal of Central Banking*, 10 (1): 95–135.

Amiti, M., and Weinstein, D.E. 2011. "Exports and Financial Shocks." *Quarterly Journal of Economics*, 126 (4): 1841–1877.

Amiti, M., and Weinstein, D.E. 2013. *How Much Do Bank Shocks Affect Investment? Evidence from Matched Bank-Firm Loan Data*. National Bureau of Economic Research, Cambridge MA.

Atanasova, C.V., and Wilson, N. 2004. "Disequilibrium in the UK Corporate Loan Market." *Journal of Banking and Finance*, 28 (3): 595–614.

Bae, K.-H., Kang, J.-K., and Lim, C.-W. 2002. "The Value of Durable Bank Relationships: Evidence from Korean Banking Shocks." *Journal of Financial Economics,* 64 (2): 181–214.

Balduzzi, P., Brancati, E., and Schiantarelli, F. 2014. *Financial Markets, Banks' Cost of Funding, and Firms' Decisions: Lessons from Two Crises*. Boston College, Boston.

Bassett, W.F., Chosak, M.B., Driscoll, J.C., and Zakrajšek, E. 2014. "Changes in Bank Lending Standards and the Macroeconomy." *Journal of Monetary Economics,* 62: 23–40.

Basten, C., and Koch, C. 2015. *Higher Bank Capital Requirements and Mortgage Pricing: Evidence from the Countercyclical Capital Buffer (CCB)*. Bank for International Settlements, Basel.

Becker, B., and Ivashina, V. 2014. "Cyclicality of Credit Supply: Firm Level Evidence." *Journal of Monetary Economics*, 62: 76–93.

Behn, M., Haselmann, R., and Vig, V. 2014. *The Limits of Model-Based Regulation*. Goethe University, Frankfurt.

Bentolila, S., Jansen, M., Jiménez, G., and Ruano, S. 2013. *When Credit Dries Up: Job Losses in the Great Recession*. Institute for the Study of Labor, Bonn.

Berg, G., and Schrader, J. 2012. "Access to Credit, Natural Disasters, and Relationship Lending." *Journal of Financial Intermediation*, 21 (4): 549–568.

Bernanke, B.S. 1983. "Nonmonetary Effects of the Financial Crisis in the Propagation of the Great Depression." *American Economic Review*, 73 (3): 257–276.

Bernanke, B.S., and Blinder, A.S. 1988. "Credit, Money and Aggregate Demand." *American Economic Review*, 78 (2): 435–439.

Bernanke, B.S., and Blinder, A.S. 1992. "The Federal Funds Rate and the Channels of Monetary Transmission." *American Economic Review,* 82 (4): 901–921.

Bernanke, B., and Gertler, M. 1989. "Agency Costs, Net Worth, and Business Fluctuations." *American Economic Review*, 79 (1): 14–31.

Bernanke, B.S., and Gertler, M. 1995. "Inside the Black Box: The Credit Channel of Monetary Policy Transmission." *Journal of Economic Perspectives*, 9 (4): 27–48.

Bernanke, B.S., Gertler, M., and Gilchrist, S. 1996. "The Financial Accelerator and the Flight to Quality." *Review of Economics and Statistics,* 78 (1): 1–15.

Bofondi, M., Carpinelli, L., and Sette, E. 2014. *Credit Supply During a Sovereign Debt Crisis*. Bank of Italy, Rome.

Bonaccorsi di Patti, E., and Sette, E. 2012. *Bank Balance Sheets and the Transmission of Financial Shocks to Borrowers: Evidence from the 2007–2008 Crisis*. Bank of Italy, Rome.

Borio, C., and Zhu, H., 2012. "Capital Regulation, Risk-Taking and Monetary Policy: A Missing Link in the Transmission Mechanism." *Journal of Financial Stability*, 8 (4): 236–251.

Brownlees, C., and Engle, R. 2015. *SRISK: A Conditional Capital Shortfall Index for Systemic Risk Measurement*. Universitat Pompeu Fabra, Barcelona.

Brunnermeier, M.K., Dong, G., and Palia, D. 2012. *Banks' Non-Interest Income and Systemic Risk*. Princeton University, Princeton NJ.

Bubb, R., and Kaufman, A. 2014. "Securitization and Moral Hazard: Evidence from Credit Score Cutoff Rules." *Journal of Monetary Economics*, 63: 1–18.

Buch, C.M., Eickmeier, S., and Prieto, E. 2014. "In Search for Yield? Survey-based Evidence on Bank Risk Taking." *Journal of Economic Dynamics and Control,* 43: 12–30.

Carbó-Valverde, S., Degryse, H., and Rodriguez-Fernandez, F. 2012. *Lending Relationships and Credit Rationing: The Impact of Securitization*. Centre for Economic Policy Research, London.

Carbó-Valverde, S., Rodriguez-Fernandez, F., and Udell, G.F. 2013. *Trade Credit, the Financial Crisis, and Firm Access to Finance*. Bangor University, Bangor.

Carvalho, D., Ferreira, M.A., and Matos, P. 2013. "Lending Relationships and the Effect of Bank Distress: Evidence from the 2007–2008 Financial Crisis." *Journal of Financial and Quantitative Analysis*, Forthcoming.

Cetorelli, N., and Goldberg, L.S. 2011. "Global Banks and International Shock Transmission: Evidence from the Crisis." *IMF Economic Review,* 59 (1): 41–76.

Chava, S., and Purnanandam, A., 2011. "The Effect of Banking Crisis on Bank-Dependent Borrowers." *Journal of Financial Economics,* 99 (1): 116–135.

Chiuri, M.C., Ferri, G., and Majnoni, G. 2002. "The Macroeconomic Impact of Bank Capital Requirements in Emerging Economies: Past Evidence to Assess the Future." *Journal of Banking and Finance,* 26 (5): 881–904.

Chodorow-Reich, G. 2014. "The Employment Effects of Credit Market Disruptions: Firm-level Evidence from the 2008–2009 Financial Crisis." *Quarterly Journal of Economics*, 129 (1): 1–59.

Ciccarelli, M., Maddaloni, A., and Peydró, J.-L. 2014. "Trusting the Bankers: A New Look at the Credit Channel of Monetary Policy." *Review of Economic Dynamics*, Forthcoming.

Cingano, F., Manaresi, F., and Sette, E. 2013. *Does Credit Crunch Investments Down? New Evidence on the Real Effects of the Bank Lending Channel*. Università Politecnica delle Marche, Ancona.

De Haas, R., and Van Horen, N., 2013. "Running for the Exit? International Bank Lending During a Financial Crisis." *Review of Financial Studies*, 26 (1): 244–285.

De Nicolò, G., Dell'Ariccia, G., Laeven, L., and Valencia, F. 2010. *Monetary Policy and Bank Risk Taking*. International Monetary Fund, Washington DC.

Del Giovane, P., Eramo, G., and Nobili, A. 2011. "Disentangling Demand and Supply in Credit Developments: A Survey-Based Analysis for Italy." *Journal of Banking and Finance,* 35(10): 2719–2732.

Del Giovane, P., Nobili, A., and Signoretti, F.M. 2013. *Supply Tightening or Lack of Demand? An Analysis of Credit Developments During the Lehman Brothers and the Sovereign Debt Crises*. Bank of Italy, Rome.

Del Prete, S., and Federico, S. 2014. *Trade and Finance: Is There More Than Just "Trade Finance"? Evidence from Matched Bank–Firm Data*. Bank of Italy, Rome.

Delis, M.D., and Kouretas, G.P. 2011. "Interest Rates and Bank Risk-Taking." *Journal of Banking and Finance*, 35 (4): 840–855.

Dell'Ariccia, G., Laeven, L., and Suarez, G., 2013. *Bank Leverage and Monetary Policy's Risk-Taking Channel: Evidence from the United States*. International Monetary Fund, Washington DC.

Dell'Ariccia, G., Laeven, L., and Marquez, R., 2014. "Real Interest Rates, Leverage, and Bank Risk-Taking." *Journal of Economic Theory*, 149: 65–99.

Demiroglu, C., James, C., and Kizilaslan, A. 2012. "Bank Lending Standards and Access to Lines of Credit." *Journal of Money, Credit and Banking*, 44 (6): 1063–1089.

Fabi, F., Laviola, S., and Reedtz, P.M. 2005. "The New Capital Accord and Banks' Lending Decisions." *Journal of Financial Stability,* 1 (4): 501–521.

Fidrmuc, J., and Hainz, C. 2013. "The Effect of Banking Regulation on Cross-Border Lending." *Journal of Banking and Finance,* 37 (5): 1310–1322.

Fraisse, H., Lé, M., and Thesmar, D. 2015. *The Real Effects of Bank Capital Requirements*. Banque de France, Paris.

Francis, W.B., and Osborne, M. 2012. "Capital Requirements and Bank Behavior in the UK: Are There Lessons for International Capital Standards?" *Journal of Banking and Finance,* 36 (3): 803–816.

Gaggl, P., and Valderrama, M.T. 2010. "Does a Low Interest Rate Environment Affect Risk Taking in Austria?" *Monetary Policy and the Economy of the Oesterreichische Nationalbank*, Q4: 32–48.

Gan, J. 2007. "The Real Effects of Asset Market Bubbles: Loan- and Firm-Level Evidence of a Lending Channel." *Review of Financial Studies*, 20 (6): 1941–1973.

Gertler, M., and Gilchrist, S. 1994. "Monetary Policy, Business Cycles, and the Behavior of Small Manufacturing Firms." *Quarterly Journal of Economics*, 109 (2): 309–340.

Grosse, S., and Schumann, E. 2014. "Cyclical Behavior of German Banks' Capital Resources and the Countercyclical Buffer of Basel III." *European Journal of Political Economy,* 34: S40–44.

Hempell, H.S., and Kok Sørensen, C. 2010. *The Impact of Supply Constraints on Bank Lending in the Euro Area – Crisis Induced Crunching?* European Central Bank, Frankfurt.

Hristov, N., Hülsewig, O., and Wollmershäuser, T. 2012. "Loan Supply Shocks during the Financial Crisis: Evidence for the Euro area." *Journal of International Money and Finance,* 31 (3): 569–592.

Ioannidou, V., Ongena, S., and Peydró, J.-L. 2014. "Monetary Policy, Risk-Taking and Pricing: Evidence from a Quasi-Natural Experiment." *Review of Finance*, 19 (1): 95–144.

Iyer, R., Peydró, J.-L., da-Rocha-Lopes, S., and Schoar, A. 2014. "Interbank Liquidity Crunch and the Firm Credit Crunch: Evidence from the 2007–2009 Crisis." *Review of Financial Studies,* 27 (1): 347–372.

Jackson, P., Furfine, C., Hans, G., Hancock, D., Jones, D., Perraudin, W., Radecki, L., and Yoneyama, M. 1999. *Capital Requirements and Bank Behaviour: The Impact of the Basle Accord*. Bank for International Settlements, Basel.

Jakovljević, S., Degryse, H., and Ongena, S. 2015. "A Review of Empirical Research on the Design and Impact of Regulation in the Banking Sector." *Annual Review of Financial Economics*, 7: Forthcoming.

Jiménez, G., Mian, A., Peydró, J.-L., and Saurina, J. 2011. *Local versus Aggregate Lending Channels: The Effects of Securitization on Corporate Credit Supply*. Banco de España, Madrid.

Jiménez, G., Ongena, S., Peydró, J.-L., and Saurina, J. 2012. "Credit Supply and Monetary Policy: Identifying the Bank Balance-Sheet Channel with Loan Applications." *American Economic Review*, 102 (5): 2301–2326.

Jiménez, G., Ongena, S., Peydró, J.-L., and Saurina, J. 2014a. "Hazardous Times for Monetary Policy: What Do Twenty-Three Million Bank Loans Say About the Effects of Monetary Policy on Credit Risk-Taking?" *Econometrica*, 82 (2): 463–505.

Jiménez, G., Ongena, S., Peydró, J.-L., and Saurina, J. 2014b. *Macroprudential Policy, Countercyclical Bank Capital Buffers and Credit Supply: Evidence from the Spanish Dynamic Provisioning Experiments*. Bank of Spain, Madrid.

Jokipii, T., and Milne, A. 2008. "The Cyclical Behaviour of European Bank Capital Buffers." *Journal of Banking and Finance*, 32 (8): 1440–1451.

Kalemli-Ozcan, S., Kamil, H., and Villegas-Sanchez, C. 2015. *What Hinders Investment in the Aftermath of Financial Crises: Insolvent Firms or Illiquid Banks?* University of Maryland, College Park.

Kashyap, A.K., Lamont, O.A., and Stein, J.C. 1994. "Credit Conditions and the Cyclical Behavior of Inventories." *Quarterly Journal of Economics,* 109 (3): 565–592.

Kashyap, A.K., and Stein, J.C. 1995. "The Impact of Monetary Policy on Bank Balance Sheets." *Carnegie-Rochester Conference Series on Public Policy,* 42: 151–195.

Kashyap, A.K., and Stein, J.C. 2000. "What Do a Million Observations on Banks Say about the Transmission of Monetary Policy?" *American Economic Review*, 90 (3): 407–428.

Kashyap, A.K., Stein, J.C., and Wilcox, D.W. 1993. "Monetary Policy and Credit Conditions: Evidence from the Composition of External Finance." *American Economic Review*, 83 (1): 78–98.

Keys, B.J., Mukherjee, T., Seru, A., and Vig, V. 2010. "Did Securitization Lead to Lax Screening: Evidence from Subprime Loans." *Quarterly Journal of Economics*, 125 (1): 307–362.

Khwaja, A.I., and Mian, A. 2008. "Tracing the Impact of Bank Liquidity Shocks: Evidence from an Emerging Market." *American Economic Review,* 98 (4): 1413–1442.

Kremp, E., and Sevestre, P. 2013. "Did the Crisis Induce Credit Rationing for French SMEs?" *Journal of Banking and Finance,* 37 (10): 3757–3772.

Laeven, L., Ratnovski, L., and Tong, H. 2014. *Bank Size, Capital, and Systemic Risk: Some International Evidence*. International Monetary Fund, Washington DC.

Loutskina, E. 2011. "The Role of Securitization in Bank Liquidity and Funding Management." *Journal of Financial Economics,* 100 (3): 663–684.

Lown, C.S., Morgan, D.P., and Rohatgi, S. 2000. "Listening to Loan Officers: The Impact of Commercial Credit Standards on Lending and Output." *Federal Reserve Bank of New York Economic Policy Review,* 6: 1–16.

Lown, C.S., and Morgan, D.P. 2002. "Credit Effects in the Monetary Mechanism." *Federal Reserve Bank of New York Economic Policy Review*, 8: 217–235.

Lown, C., and Morgan, D.P., 2006. "The Credit Cycle and the Business Cycle: New Findings Using the Loan Officer Opinion Survey." *Journal of Money, Credit and Banking,* 38 (6): 1575–1597.

Maddaloni, A., and Peydró, J.-L. 2011. "Bank Risk-Taking, Securitization, Supervision, and Low Interest Rates: Evidence from the Euro-area and the U.S. Lending Standards." *Review of Financial Studies*, 24 (6): 2121–2165.

Mutu, S., and Ongena, S., 2015. *The Impact of Policy Interventions on Systemic Risk across Banks*. University of Zurich, Zurich.

Ogawa, K., and Suzuki, K. 2000. "Demand for Bank Loans and Investment under Borrowing Constraints: A Panel Study of Japanese Firm Data." *Journal of the Japanese and International Economies*, 14: 1–21.

Oliner, S.D., and Rudebusch, G.D. 1996. "Is there a Broad Credit Channel for Monetary Policy?" *Federal Reserve Bank of San Francisco Economic Review*, 1: 3–13.

Ongena, S., Peydró, J.L., and van Horen, N. 2015. *Shocks Abroad, Pain at Home? Bank-Firm Level Evidence on the International Transmission of Financial Shocks*. De Nederlandsche Bank, Amsterdam.

Ongena, S., Popov, A., and Udell, G.F. 2013. "'When the Cat's Away the Mice Will Play': Does Regulation at Home Affect Bank Risk-Taking Abroad?" *Journal of Financial Economics*, 108 (3): 727–750.

Ongena, S., Smith, D.C., and Michalsen, D. 2003. "Firms and their Distressed Banks: Lessons from the Norwegian Banking Crisis." *Journal of Financial Economics,* 67 (1): 81–112.

Ongena, S., Tümer-Alkan, G., and von Westernhagen, N. 2015. *Do Exposures to Sagging Real Estate, Subprime or Conduits Abroad Lead to Contraction and Flight to Quality in Bank Lending at Home?* Deutsche Bundesbank, Frankfurt.

Paravisini, D., Rappoport, V., Schnabl, P., and Wolfenzon, D. 2015. "Dissecting the Effect of Credit Supply on Trade: Evidence from Matched Credit-Export Data." *Review of Economic Studies*, 82 (1): 333–359.

Peek, J., and Rosengren, E. 1995. "The Capital Crunch: Neither a Borrower nor a Lender Be." *Journal of Money, Credit, and Banking*, 27 (3): 625–638.

Peek, J., and Rosengren, E.S. 1997. "The International Transmission of Financial Shocks: The Case of Japan." *American Economic Review,* 87 (4): 495–505.

Peek, J., and Rosengren, E.S. 2000. "Collateral Damage: Effects of the Japanese Bank Crisis on Real Activity in the United States." *American Economic Review*, 90 (1): 30–45.

Popov, A., and Udell, G.F. 2012. "Cross-Border Banking, Credit Access, and the Financial Crisis." *Journal of International Economics*, 87 (1): 147–161.

Presbitero, A.F., Udell, G.F., and Zazzaro, A. 2014. "The Home Bias and the Credit Crunch: A Regional Perspective." *Journal of Money, Credit and Banking*, 46 (s1): 53–85.

Puri, M., Rocholl, J., and Steffen, S. 2011. "Global Retail Lending in the Aftermath of the US Financial Crisis: Distinguishing between Supply and Demand Effects." *Journal of Financial Economics,* 100 (3): 556–578.

Rajan, R.G. 2005. *Has Financial Development Made the World Riskier?* National Bureau of Economic Research, Cambridge MA.

Romer, C.D., and Romer, D.H. 1990. "New Evidence on the Monetary Transmission Mechanism." *Brookings Papers on Economic Activity*, 1: 149–213.

Schnabl, P. 2012. "The International Transmission of Bank Liquidity Shocks: Evidence from an Emerging Market." *Journal of Finance*, 67 (3): 897–932.

Shim, J. 2013. "Bank Capital Buffer and Portfolio Risk: The Influence of Business Cycle and Revenue Diversification." *Journal of Banking and Finance*, 37 (3): 761–772.

Slovin, M.B., Sushka, M.E., and Polonchek, J.A. 1993. "The Value of Bank Durability: Borrowers as Bank Stakeholders." *Journal of Finance*, 48 (1): 247–266.

Tabak, B.M., Noronha, A.C.B.T.F., and Cajueiro, D.O. 2011. *Bank Capital Buffers, Lending Growth and Economic Cycle: Empirical Evidence for Brazil*. Second BIS CCA Conference "Monetary Policy, Financial Stability and the Business Cycle", Ottawa, Canada.

Tarashev, N., Borio, C., and Tsatsaronis, K., 2010. *Attributing Systemic Risk to Individual Institutions*. Bank for International Settlements, Basle.

van der Veer, K., and Hoeberichts, M. 2013. *The Level Effect of Bank Lending Standards on Business Lending*. De Nederlandsche Bank, Amsterdam.

VanHoose, D.D. 2008. "Bank Capital Regulation, Economic Stability, and Monetary Policy: What Does the Academic Literature Tell Us?" *Atlantic Economic Journal*, 36 (1): 1–14.

Woo, D. 2003. "In Search of 'Capital Crunch': Supply Factors behind the Credit Slowdown in Japan." *Journal of Money, Credit and Banking,* 35 (6): 1019–1038.

Zia, B.H. 2008. "Export Incentives, Financial Constraints, and the (Mis)allocation of Credit: Micro-level Evidence from Subsidized Export Loans." *Journal of Financial Economics*, 87 (2): 498–527.

3
Market Discipline, Public Disclosure and Financial Stability

Rhiannon Sowerbutts and Peter Zimmerman

1 Introduction

Inadequate disclosure by commercial banks has been cited as a contributing factor to the financial crisis. Banks did not report enough information about the assets they were holding or the risks that they were exposed to, and inadequate disclosure meant that investors were less able to judge risks to a bank's solvency than bank insiders, such as managers. Investors did not demand sufficient disclosure prior to the crisis. Possible reasons for this include risk illusion, or expectations that governments would be willing and able to bail out failing banks.

Increased uncertainty aversion during a time of systemic stress led to investors withdrawing funding from the most opaque banks. The lack of transparency is likely to have intensified the crisis – for example, by leading to much higher funding costs, even for relatively healthy banks. Increased disclosure can help to alleviate the problem of asymmetric information between banks, who have good information about their own financial resilience, and investors that provide funding to banks, who have less information.

Better disclosure can be beneficial to financial stability in non-crisis times, too. With good information, debt investors are able to price risk more accurately and, if the incentives are right, this can act as a disciplining force on banks. As debt investors become aware of the risks that banks are taking, they are less likely to provide funding to banks that are not providing an attractive trade-off between risks and returns. This can affect the risk-taking decisions of bank managers. This market discipline mechanism empowers investors to ensure that managers are acting in their interests, and reduces the likelihood that a bank takes risks that its investors are not aware of. Therefore publishing better information may reduce the probability of future financial crises, as it can make sudden changes in investor sentiment less likely.

Public disclosure reduces information asymmetries between insiders (managers of banks) and outsiders (investors), and so means greater certainty for investors in their ability to forecast the performance of banks' debt and equity. In a perfectly functioning market, investors would demand that managers of banks disclose information about risks in order to allow those investors to correctly price the banks' liabilities. In principle, in the absence of social externalities this market discipline mechanism could make prudential regulation redundant: investors would ensure that banks do not behave in a socially harmful way by influencing management. The idea that investors may be able to effectively monitor financial institutions and constrain socially harmful risk-taking has been a cornerstone of regulatory policy for years. Basel II explicitly states that the purpose of "market discipline is to complement the minimum capital requirements (Pillar 1) and the supervisory review process (Pillar 2). The [Basel] Committee aims to encourage market discipline by developing a set of disclosure requirements which will allow market participants to assess key pieces of information on the scope of application, capital, risk exposures, risk assessment processes, and hence the capital adequacy of the institution" (Basel Committee on Banking Supervision 2006).

However, frictions exist which prevent this market discipline channel from functioning correctly. That leads to information asymmetries, a tendency for banks to become overly leveraged, and a higher probability of banking crises, all of which reduce social welfare.

Mandatory disclosure policies can – if correctly calibrated – correct for these market failures and increase social welfare. These can act as a complement to prudential regulation, allowing both market participants and regulators to take responsibility for ensuring that bank managers' incentives are aligned with those of their stakeholders, and leaving regulators to address any externalities to which stakeholders do not attend. This chapter discusses the evidence for whether investors monitor the financial institutions in which they invest and the reasons why this "monitoring channel" may break down. We conclude with a discussion of whether more information and increased market discipline is actually optimal for financial stability.

2 Modeling and measuring market discipline: testing the "monitoring channel"

Empirical studies disagree on whether private sector agents reliably engage in risk monitoring. Researchers cannot directly observe whether every agent pores over financial statements, or participates in conference calls with banks. In practice, testing for whether investors monitor a bank usually means examining whether the return that private sector agents demand is commensurate with the risk that they face.[1] Prior to the 1990s, studies generally fail to find a significant relationship between bank risk and the yields investors demand. However, subsequent studies find evidence of market discipline: Ellis and Flannery (1992), James (1991), Keeley (1990) and Flannery and Sorescu (1996) all find that high certificate of deposit rates and subordinated debt spreads reflect different measurable elements of bank risk, providing evidence for the existence of market discipline. The change in results can perhaps be attributed to the FDIC and the Federal Deposit Insurance Corporation Improvement Act, passed in 1991 following the US savings and loan crisis, which made the safety net for banks more restrictive. Sironi (2003) finds similar results for Europe: he examines subordinated debt and debentures issued in Europe from 1991–2000 and finds that the sensitivity of subordinated debt issues to measures of standalone risk (i.e., without incorporating external guarantees) increased during the 1990s.

However, as Gorton and Santomero (1990) point out, a number of the studies above suffer from a failure to take into account how investors should respond in theory to the variables that they measure. In particular, many of these studies assume that the value of subordinated debt is a monotonic function of bank risk-taking. But, as Black and Cox (1976) show, while junior debt is initially a convex function of the value of the firm, it becomes a concave function when the value of the firm is sufficiently high. Unlike senior debt, the default risk premium on subordinated debt is a *decreasing* function of the riskiness of a firm's assets when the firm is close to bankruptcy and then an *increasing function* when the bank is relatively far away.

The intuition for this result is fairly simple: it arises from the fact that the deadweight cost of bankruptcy for banks is large. James (1991) estimates that direct expenses associated with bank failures are on average 10 percent of assets, with an average loss of 30 percent. As subordinated debt and equity tend to comprise a smaller proportion of banks' balance sheets than this, subordinated debt will receive a payoff that is close to zero in the event of bank failure. This means that, when a bank is close to failure, then subordinated debt has a risk-reward payoff similar to equity – i.e., it is initially zero, and its value increases with the risk-taking of the bank. But, when the probability of bank failure is low, subordinated debt behaves more like senior debt, and its value should decrease with the risk-taking of the bank. However, the studies mentioned above tend to assume that the default risk premium is an increasing function of riskiness; this means that assuming a linear model at a time when a bank is close to failure (and so the premium is actually decreasing in risk) will lead to an underestimate of the extent of market discipline.

All the above studies essentially focus on the change in the rate of return investors demand for bearing increased risk. The next section discusses how to measure this.

2.1 Finding empirical evidence of market discipline

Market discipline may be more evident in the markets for some instruments than others. For example, premiums on senior debt instruments may not be sufficiently sensitive to credit risk. A good candidate should be risk-sensitive, have reliable price data, and have a long residual maturity (so that investors cannot simply respond to credit risk by allowing the instrument to mature). Collateralized or government-guaranteed debt is clearly unsuitable for these purposes. As noted earlier most studies examine subordinated debt or large certificates of deposit. Several more recent studies use spreads on credit default swaps, which was not a market that existed in the early 1990s.

Market liquidity is crucial for price data to be reliable. For this reason most US studies are able to use secondary market data, but for European banks, liquidity in the secondary market for subordinated debt is often poor. Therefore, any econometric study of secondary market data should employ liquidity controls. This is a difficult area, as the section below discusses, because traditional measures of liquidity are heavily influenced by information asymmetries.

One solution may be to use data on primary issuance, as it reflects an updated assessment of risk premiums by investors purchasing the bonds. However, the decision of whether or not to issue in any given time period may be a form of market discipline in itself: for a risky bank, the required premium may be high enough to induce it not to issue subordinated debt, but instead to issue another less-sensitive instrument, or to delay issuing debt at all. It may be sensible to run a probit/logit model to test whether the decision to issue subordinated debt is affected by bank-specific risks. In any case, primary issuance is not usually a frequent event for any individual bank, so a large time series would be required to avoid small-sample problems in a fixed effects regression.

Controls specific to the particular instrument are needed. Time to maturity, the seniority of the issue and liquidity of the bond are all obvious candidates. But there are controls which are particularly related to information and market discipline. The most notable of these is probably *issue size*. When information is costly to analyze and monitor, major buyers of subordinated debt may prefer to specialize. If so, they would purchase large amounts of debt of a small number of firms.

2.2 Measuring bank risk

One of the most important challenges in determining market discipline is how to measure bank risk. For investors to exert discipline, they must be able to observe bank risk. Measures of bank risk can be broadly categorized as accounting-based measures (those based on firms' published balance sheet information), ratings-based measures (based on the assessment of credit ratings agencies or other delegated monitors) and market-based measures (i.e., those based on the prices of traded instruments).

2.2.1 Accounting-based measures

The Z-score developed by Boyd and Graham (1988) is an accounting-based measure of bank's distance to default – that is, the number of standard deviations that a bank's return on assets can fall by before it becomes insolvent. It is calculated as the sum of return on assets and the equity-to-asset (leverage) ratio divided by the standard deviation of the return on assets, usually measured over four quarters to allow sufficient variation in the return on assets. This measure encompasses a number of popular accounting measures such as the standard deviation of return on equity/assets and leverage. Caution is needed for interpretation. A higher return on assets could reflect higher risk-taking, but it may also represent greater efficiency, making default less likely.

Other popular accounting measures of credit risk include the proportion of non-performing loans and concentration of lending in a particular sector. However, modeling approaches which focus solely on credit risk do not capture important elements of bank risk-taking, such as liquidity and trading risk. Since the 1999 repeal of the US Glass-Steagall act – which separated trading and lending activities – trading and wholesale funding have become an important part of the activities of commercial banks, making it more important for the recent literature to focus on these risks. Liquidity risk can be captured in a number of different ways: past papers have tended to focus on the liability side and used some kind of ratio of short-term debt to total debt. More recent papers such as Sironi (2003) consider liquidity on the asset side of the balance sheet too.

2.2.2 Ratings-based measures

Credit ratings can be considered an amalgamation of all the risk factors above in a summary statistic: the credit rating. These have some advantages over accounting or market-based measures of risk in that they are more standardized allowing for better cross-country comparisons. For example, the definition of non-performing loans varies considerably across countries and this can be difficult for an individual investor to analyze. Credit ratings aim to rate "through the cycle" meaning that they should be forward-looking and take into account macroeconomic conditions. Crucially, credit rating agencies are delegated monitors: the ratings are public information and free to acquire. The downside is that these are the subjective opinion of a rating agency, and ratings are slow to be updated in response to events and emerging risks.[2]

2.2.3 Market-based measures

Market-based measures use observable and timely information from market prices, rather than relying on accounting information or delegated monitors.

For example, "distance to default" is a market measure of credit risk analogous to the Z-score mentioned above. It is based on the seminal work of Merton (1974), which treats the equity value of the firm as a call option on the firm's assets. Distance to default is the difference between the asset value of the firm and the face value of its debt, scaled by the standard deviation of the firm's asset value. In other words, it is a proxy for the likelihood of the bank being unable to pay its debt in future: a higher distance to default implies a lower probability of insolvency. The value and standard deviation of the firm's assets can be derived using market prices for equity in the Merton model framework. This metric is commonly used for non-financial corporates but has also been often applied to banks, particularly for the analysis of deposit insurance payouts.

2.2.4 Comparing market-based and accounting-based measures

Market-based measures contain information that is absent from accounting ratios. The data is timelier, less prone to manipulation and less targeted. By contrast, accounting data is backwards-looking and is released infrequently, with between reports gaps of at least a quarter being common practice. If the equity market is at least semi-strong form efficient, then it should contain all of the relevant data from previous publications of accounts, so it may be argued that market-based measures contain strictly more information than accounting-based measures.

This can be illustrated by comparing regulatory capital requirements – which are based on accounting measures and so are backwards-looking – with a market-based capital ratio equivalent. Figure 3.1a compares Basel II Tier 1 capital ratios for banks which did and did not fail during the period of most intense financial market distress in autumn 2008. At the time, this was the prevailing measure of regulatory capital. As can be seen, there is no discernible difference between the two groups of banks, suggesting that this measure is a poor predictor of distress.

Figure 3.1b presents, for the same banks, a market-based equivalent, namely the ratio of market capitalization (based on the contemporaneous traded share price) divided by book value of assets. As can be seen, the two sets of banks can be clearly distinguished under this measure, which is a much better predictor of bank failure. As Haldane (2011) states, market-based measures offer the advantage of simplicity and transparency: "200 million separate calculations would condense to a simple sum."[3]

This is not to say that accounting-based ratios are useless. The drawbacks of Basel II even as an accounting-based measure of risk are well-documented: we find, for example, that regulatory capital ratios under Basel III have a higher correlation with market-based capital ratios, suggesting that they may do better at predicting crises. See Table 3.1 below.

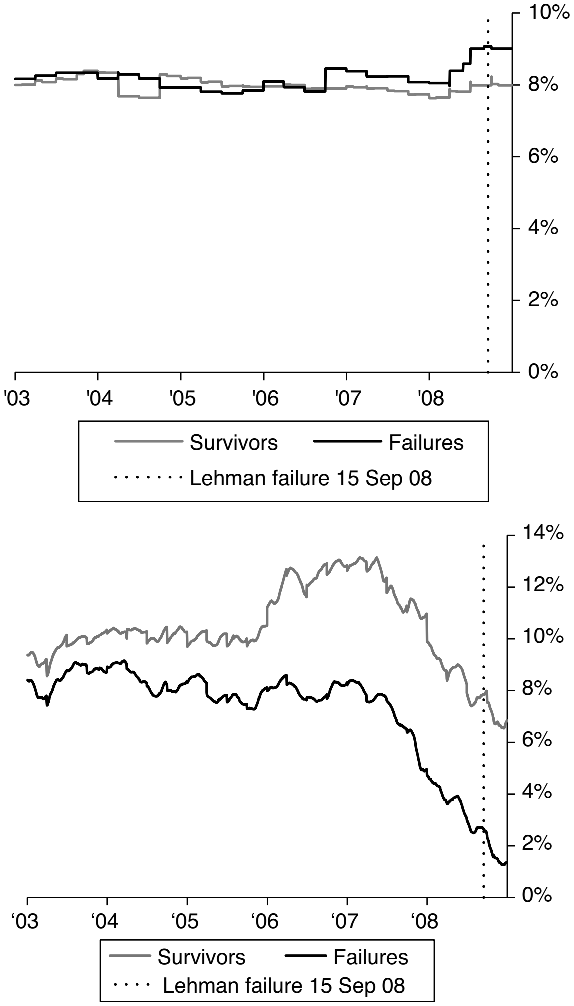


Figure 3.1 a Basel II tier 1 capital ratio as a predictor of bank distress, January 2003–December 2008; b Market-based capital ratio as a predictor of bank distress, January 2003–December 2008

Notes: (a) "Failures" are a set of major financial institutions, which in autumn 2008 either failed, required government capital or were taken over in distressed circumstances. These are RBS, HBOS, Lloyds TSB, Bradford & Bingley, Alliance & Leicester, Citigroup, Washington Mutual, Wachovia, Merrill Lynch, Freddie Mac, Fannie Mae, Goldman Sachs, ING Group, Dexia and Commerzbank. The chart shows an unweighted average for those institutions in the sample for which data are available on the given day.

(b) "Survivors" are HSBC, Barclays, Wells Fargo, JP Morgan, Santander, BNP Paribas, Deutsche Bank, Crédit Agricole, Société Générale, BBVA, Banco Popular, Banco Sabadell, Unicredit, Banca Popolare di Milano, Royal Bank of Canada, National Australia Bank, Commonwealth Bank of Australia and ANZ Banking Group. The chart shows an unweighted average for those banks in the sample for which data are available on the given day.

(c) 30-day moving average of market-based capital ratio measure.

Source: Capital IQ and authors' calculations.

Table 3.1 Correlations with market-based capital ratio

Basel II tier 1 capital ratio	Basel III core equity tier 1 capital ratio	Basel III leverage ratio
0.39	0.81	0.89

Source: Bloomberg, reported data and authors' calculations. Data are for five largest UK banks, Dec 2011–Nov 2012. This short period is selected as one in which banks may reasonably be thought to be targeting both Basel II and Basel III capital ratios.

Of course, one reason for this may be that in the recent post-crisis period, equity investors reward banks with healthy capital ratios under the new regime – we cannot be so sure that this correlation will remain strong in crisis times. Moreover, as a bank may have to enter resolution if its regulatory capital ratio falls below a certain level, it will aim to maintain a constant, healthy ratio above almost all other objectives.

2.3 Equity investors

Measures of market discipline typically relate to the response of debt investors to changes in risk. This is partly because the literature has focused on the disciplining role of debtors and the conflict between debt and equity investors. But examining the pricing of equity can shed light on the issue too. In particular, if we observe that equity investors distinguish between banks but debt investors do not, then we may be able to rule out that the failure of market discipline for creditors is due to a lack of information or an inability to process it. More bluntly, it may be that equity investors monitor the bank, while debt investors fail to do so.

However, caution should be drawn against jumping to this conclusion. An alternative explanation is that in the econometric analysis carried out in the literature, too much is asked of debt investors to distinguish between banks on the basis of the variables used, especially when measures of risk are used that are based on potentially manipulated or targeted accounting measures (such as regulatory capital ratios as explored above).

Moreover, the payoffs of equity and debt – and their sensitivity to underlying risk – are very different, and vary with the state of the world. For equity investors, a change in risk in any state of the world in which the bank is solvent (or close to being so) will affect their payoff. But, by contrast, the sensitivity of debt to risk is higher in states of the world where the bank is insolvent or close to being so. This means that debt investors may require different information to equity investors, and it may be harder to collect, especially as firms' disclosure policies are more likely to be driven by shareholders' rather than creditors'

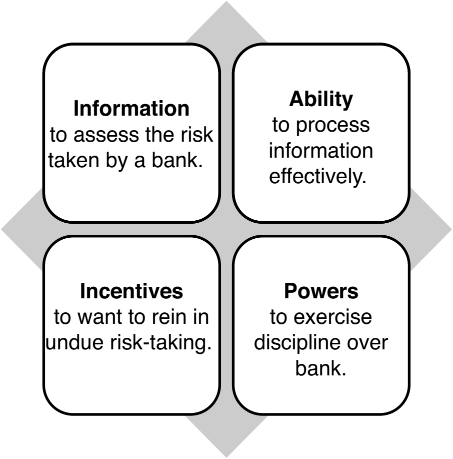

Information
to assess the risk taken by a bank.
Ability
to process information effectively.
Incentives
to want to rein in undue risk-taking.
Powers
to exercise discipline over bank.

Figure 3.2 The four requisites for effective market discipline
Source: Sowerbutts et al. (2013).

preferences. The reasons why market discipline may break down is the subject of the next section of this chapter.

3 Why market discipline can break down

Our discussion so far has focused on the issue of whether or not investors respond to the risks that banks are taking. The literature finds evidence that investors do not effectively impose market discipline, neither by monitoring bank risk nor by influencing management.

This section focuses on the reasons *why* market discipline can break down. Crockett (2001) identifies four requisites for effective market discipline, which are illustrated in Figure 3.2 above. Debt investors need to have: sufficient information to understand the risks that banks are taking; the ability to process this information; powers to discipline banks to rein in risk-taking where necessary; and incentives to exercise these powers.

3.1 Do investors have the information that they need?

Sowerbutts et al. (2013) introduce a quantitative framework to assess the first of these channels. Their metric assesses whether investors have sufficient information to understand the risks that banks are facing in a number of areas: funding risk; group structures; asset valuation; intra-annual information and financial interconnections. These contrast with the measures of risk mentioned in the previous section, which mainly focus on credit risk.

However, the financial crisis revealed that disclosure in these other non-credit areas had been insufficient prior to the crisis, and that investors had failed to demand that banks disclose more (Bank of England 2009). Therefore measuring improvements in these areas in the post-crisis period is a useful way of tracking whether disclosure has improved. For this reason, the index measures disclosure over and above minimum international regulatory standards.

The index in Sowerbutts et al. scores a bank between zero and one for each indicator, depending on whether the relevant information was disclosed in a public annual report. The figures below show that on a global level there has been a broad improvement in disclosures since the crisis. Figures 3.3a, b and c show the average disclosure scores for three of these categories over the period 2000–2012. Each line shows the average for the group of banks in that jurisdiction. There is an upward trend in all three categories, though progress varies between jurisdictions.

This kind of quantitative index cannot capture qualitative or subjective information such as clarity of exposition in banks' disclosures or standardization and comparability of reporting. Even so, it is very labor-intensive to produce. The US Securities and Exchange Commission requires standardized templates for financial reporting (10-Q and 10-K reports), but in general this is not the case in most other countries, where lack of standardization between reporting makes comparability harder. Accounting standards vary between countries and are often principles-based. This means that management must use its judgment in providing reliable and relevant information, and this could lead to substantial variation between banks. To the extent that market discipline is effective, investors may wish to encourage management to standardize reporting between banks and across time, to make direct comparability easier.

3.2 Shedding light on bank opacity

Opacity can be characterized as three nested cases: some outsiders (i.e., investors) are informed; only insiders (managers) are informed; or the business is fundamentally unknowable, even by managers. Figure 3.4 illustrates.

Each of these cases can be analyzed and measured separately, and each leads to different predictions for asset prices. This subtlety may explain some of the apparently conflicting results which exist in the literature.

3.2.1 Measuring asymmetric information between managers and investors

If outsiders are unable to completely observe the firm's actions, then managers will have some ability to capture cash flows for their private benefit. However, if agents are aware of this, they can increase the return they demand with the expected value of the missing information. This has important implications for the cost of raising equity. In the "pecking order theory" model introduced by Myers and Majluf (1984), managers have more information than outside

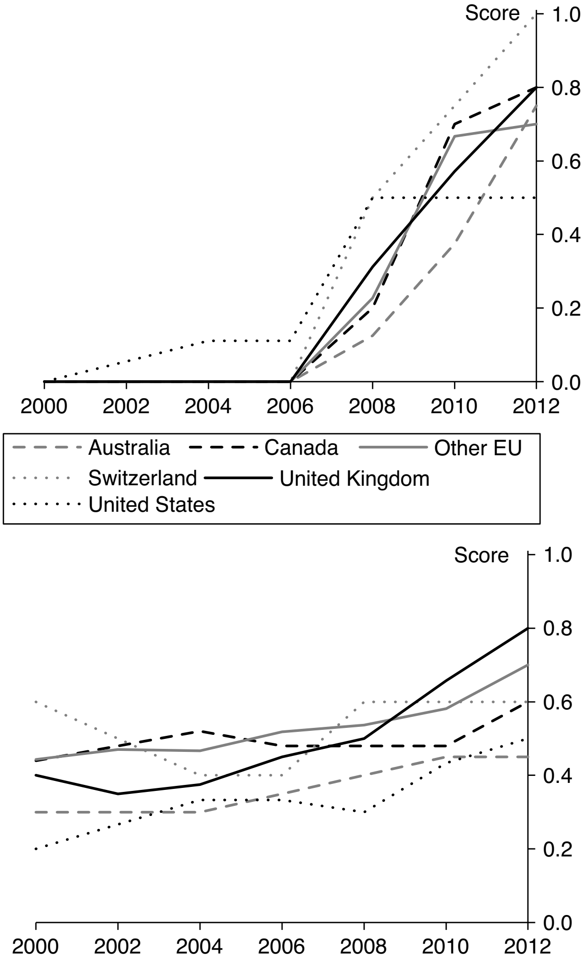


Figure 3.3 a Valuation category scores; b Funding risk category scores; c Financial interconnections category scores

Notes: (a) Each category assigns a score between 0 and 1 for a bank based on whether or not detailed quantitative disclosure takes place. The scores are measured for a panel of 50 banks – these scores show the progress made by jurisdiction.

(b) "Valuation" score assesses disclosure of valuation methodology and sensitivity to the underlying assumptions.

(c) "Funding risk" score assesses disclosure of funding breakdown across five different metrics: by type, maturity, currency, asset encumbrance, and a stress ratio measure.

(d) "Financial interconnections" assesses disclosure of exposures to other banks and off-balance sheet entities, as well as implicit support.

Source: Sowerbutts et al. (2013).

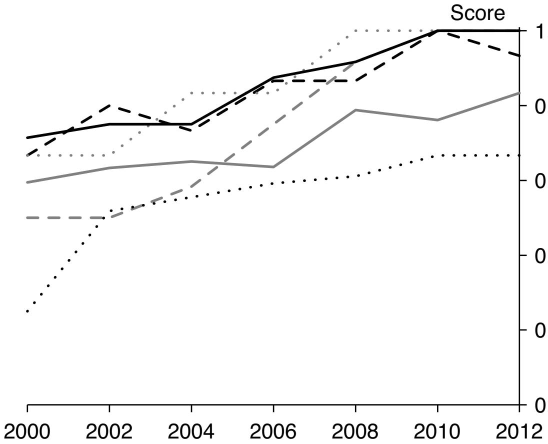


Figure 3.3 (*Continued*)

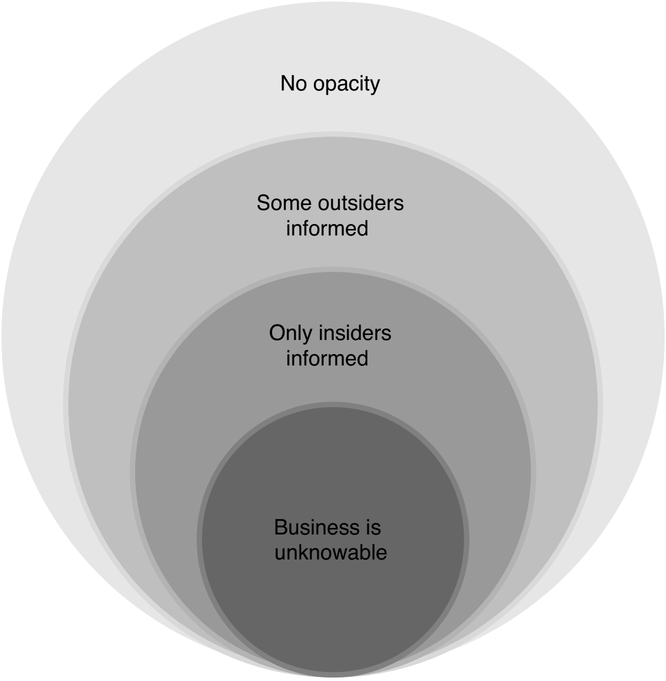


Figure 3.4 Opacity in firms

investors, and so investors perceive issuance of equity as a negative signal of managers' expectations of future firm value. This makes raising equity more expensive than issuing debt or using internal funds to finance projects. An implication of this is that more opaque firms – for example, banks – will be more leveraged than more transparent firms.

However, there are a number of other factors that contribute to capital structure decisions, making this theory challenging to test empirically, particularly for large and complex firms such as banks. Indeed, most tests of Modigliani-Miller's capital structure invariance hypothesis – which postulates that a firm's total funding costs should be independent of capital structure – exclude banks and other financial firms: they tend to have more complex capital structures than other types of firm. Jin and Myers (2006) develop an extension to Myers (2000) in which investors receive news that is a combination of firm-specific information and macroeconomic or industry information. The predictions are fairly clear: firms with more managerial inside information will have equity returns which are less likely to reflect firm-specific information and instead equity returns will be more likely to reflect market (and perhaps industry) information. Several studies examine the relationship between this type of information symmetry and the goodness of fit (R-squared) from asset pricing regressions. Haggard and Howe (2012) test this prediction by comparing banks to non-financial firms with similar equity market characteristics. Their results suggest that banks are more subject to this form of insider-outsider information asymmetry than other types of firm.

3.2.2 Measuring asymmetric information between investors

Easley and O'Hara (2004) develop a theoretical model of informed and uninformed investors and show that investors demand a higher return for holding assets with greater private information. This is because private information increases the risk to uninformed investors of holding the asset, and this risk cannot be diversified away.

Aspects of market microstructure are frequently used to analyze asymmetric information between investors. If all investors know all the information about an asset – and agree that they do – then it will trade with a small bid-ask spread. But, when some investors have private information, bid-ask spreads will increase as market makers seek to protect themselves against trading with informed traders. The greater the proportion of informed traders, the less likely price changes are to be reversed (Kyle 1985). But predictions on volume are unclear. If *no* investor knows an asset's fundamental value then it can be very liquid (Dang et al. 2013), as market makers have no concerns about information asymmetries. But as soon as some trader has some private information about the asset value, then this market can break down as uninformed investors are not willing to hold the asset. The market microstructure literature generally decomposes the quoted bid-ask spread into three components: order-processing costs, inventory-holding costs, and adverse-selection costs.

Flannery et al. (2004) is an important paper in this literature, using these market microstructure measures of bank's equity to measure opacity. Opacity is defined to mean that some investors cannot value the asset very accurately but (perhaps) insiders or informed traders can; by this definition a more opaque asset would have a bigger bid-ask spread. The authors also examine data on analyst earnings forecasts, measuring both accuracy and dispersion: they conclude that "banking assets are not unusually opaque; they are simply boring." They also find that forecast dispersion for banks is virtually indistinguishable from non-banks, and that the median forecast errors are smaller for non-banks – although the latter result could be due to banks being better able than other firms to "manage" their earnings to meet analysts' expectations. Large bank holding companies (BHCs) are found to have similar trading properties to their matched non-financial firms, suggesting that they are as transparent as similar large non-financial firms.[4] But smaller BHCs trade much less frequently than comparable non-banks, despite having similar bid-ask spreads. They also have lower return volatilities and are more easily forecastable relative to comparable non-financial firms, suggesting that banks are not especially opaque.

Flannery et al. (2013) repeat the same exercise but over a longer time period, which incorporates the global financial crisis and the failure of LTCM in 1998. They find that, although banks are no more opaque than their non-financial counterparts pre-crisis, during crisis times both the spreads and price impacts of BHC stocks are significantly higher than those of non-banks. As the authors note, "The general pattern of time-varying relative bank opacity is troubling, since it suggests a reduction in bank stability during crisis periods, even beyond the obvious deterioration in bank balance sheet values."

3.2.3 Unknowable business models and information uncertainty

Morgan (2002) uses a very simple model to capture uncertainty, looking at disagreement among credit rating agencies, who are considered to be insiders with access to private information about the firm. Morgan considers the hypothesis that, if risk is harder to observe in banks than non-banks, then rating agencies should disagree more over ratings to a greater extent. This means that opacity can be proxied using statistics such as the average difference between ratings, their correlation, and the percentage of issues where there is disagreement between agencies. For a sample of bonds issued by firms between 1983 and 1993, Morgan finds that disagreement is greater for bank bonds than non-bank bonds. Interestingly he shows that rating agencies disagreed more about banks after 1986, which he attributes to the demise of the "too big to fail" safety net in the US following the collapse of Continental Illinois and subsequent regulatory reform. One weakness of this argument is that, until the global financial crisis, the rating agencies in the paper – S&P and Moody's – had not properly formalized the way that government support was factored into ratings, making it harder to test these hypothesis. The amount of disagreement is increasing in the level of loans and trading assets of banks, and in their degree of leverage. Morgan interprets these results as suggesting that it is the business of banking that makes it inherently more opaque than other industries.

Iannotta (2006) undertakes a similar analysis for European banks and finds similar results, although he also identifies construction, energy and utility and "other" as being more opaque than the banking industry. He also finds that a higher capital ratio increases the likelihood of a split rating.

3.3 Using stress tests to assess opacity

A number of papers use the recent US and EU bank stress test disclosures as a way to further unpack the sources of bank opacity. In the wake of the recent financial crisis, regulators have regularly carried out stress tests on banks and published the results in order to increase confidence in the banking system. This has created an ideal environment to study empirical informational issues. In contrast to their own disclosures, firms are unable to choose what is disclosed in the stress test. Information contained in stress test disclosures is often considered by investors to be as informative as the outcome of the test itself. In addition, investors may draw inferences from a regulator's selection of stress scenario, or the banks chosen to participate in the test.

Stress tests are concerned with a downside scenario, which is of direct interest to debt investors. This contrasts with information in annual reports, which is generally designed to be informative to shareholders. Therefore disclosure of stress test results can complement disclosure that banks voluntarily provide, giving all market participants information required to assess the risks of the instruments that they hold.

Morgan et al. (2014) examine whether the 2009 US stress tests were informative to investors. They measure information using several events around the tests and calculate cumulative abnormal equity returns. They find a significant negative relationship between abnormal returns around the release of the stress test results and the capital gap that banks were found to have; this is consistent with the view that the stress test produced information about the banks the private sector analysts did not already have. However, a recent paper by Glasserman and Tangirala (2015) show that there is some predictability in the stress test outcomes over time, and so diversity of scenario design can ensure that the tests remain meaningful for investors.

Ellahie (2013) examines the European Banking Authority stress tests in 2010 and 2011 and tests for information asymmetry across investors – using bid-ask spreads – and information uncertainty, using equity option-implied volatilities and relative CDS spreads for one- vs. five-year debt. Unfortunately, the author is unable to empirically disentangle whether the increased uncertainty is due to greater underlying volatility from the worsening sovereign credit crisis, or due to poor quality information contained in the stress test disclosures. This is something that plagues the earlier literature on stress tests, which were frequently undertaken in crisis conditions. More regular stress testing – and a more tranquil economic environment – will hopefully overcome this problem; for example in the UK stress tests will take place on an annual basis (Bank of England 2013). But this may result in the risk – at least, for research purposes – that investors will pay less attention to stress test disclosures in non-crisis times.

3.4 The big black box of banking?

The above findings suggest that banking is – to an extent – a business which is opaque and difficult to assess. Even before the crisis, banking was described as a "black box" (see, e.g., *The Economist* 2007). The nature of banking is by some definitions, opaque. One of the many functions of banks is to overcome information asymmetries and lend to borrowers who are unable to raise market finance or who may wish to use bank finance to avoid disclosing sensitive information. A bank may have advantages in being able to screen borrowers, overcome moral hazard, and negotiate in default. This has to suggestions that banks are inherently opaque. But opacity at the individual loan level does not mean that the portfolio of a bank must necessarily be opaque, nor that the payoffs of its liabilities must be. A classic and simple example of this can be found in the model of Diamond (1984), in which creditors do not monitor the bank's individual loans but do understand the bank's incentives and so are perfectly informed about the bank's portfolio.

3.5 Guarantees in the banking system

The literature on market discipline in banks almost disappears in the early 2000s. This reflects a number of factors. One reason is that policy interest declined considerably after the 1990s: while a paper by the Board of Governors of the Federal Reserve System and Secretary of the Treasury (2000) counts no fewer than 14 proposals for mandatory subordinated debt issuance made in the 1990s, interest becomes scarcer after this period until just before the global financial crisis. The importance of market discipline was cemented in Pillar 3 of Basel II, which encourages greater disclosure of a bank's risk in order to enhance market discipline.

Another factor is that interest in the literature turned away from explicit discussion of market discipline, and focused more on assessment and measurement of the implicit subsidy of banks. However, many of the techniques used are similar. A recent strand of the literature attempts to examine whether there is an implicit guarantee for banks which are expected to be bailed out by governments in the event of failure – in other words, which are "too big to fail."

Typically these papers examine whether there is a relationship between risk-taking by banks and the spread that investors demand, and then analyze how this is affected by expectations of implicit guarantees. It is important to bear in mind that, even though in many cases pricing does appear to be risk-sensitive, this does not necessarily mean that there is no guarantee. The question is: is risk priced *enough?*

Acharya et al. (2014) adapt the usual market discipline equation – i.e., a fixed effects regression of spread against credit risk – but add a number of control variables, all of which capture different measurements of risk. The authors use a number of different measures of "too big to fail" status, such as CoVaR (which measures a firm's contribution to systemic risk) and bank size. Their results suggest that large institutions have lower spreads due to implicit government support, not because they have lower risk. Siegert and Willison (2015) provide a literature review of similar papers.

Morgan and Stiroh (1999) investigate market discipline using bond spreads, ratings and bank data for bonds issued between 1993 and 1998 and find evidence that the spread on bank bonds increases as credit ratings deteriorate. However, they show that this effect is weaker for bigger and less transparent banks, pointing to possible slippage in the disciplinary mechanism for banks either considered too big to fail or too hard to understand by the bond market.

Guarantees and government support of any form can undermine market discipline as they disrupt the transmission of the risks that a bank is taking into the risks that investors actually face. This does not always mean that guarantees would increase incentives for banks to increase risk-taking: the effect will depend on the nature of the guarantee. In a case where a bank is insured against all losses, then it will certainly seek to maximize risk-taking. But support which occurs in states of the world that are not strongly correlated with the bank's risk choices – for example, lender of last resort activities in the event of systemic liquidity shortages – is less likely to distort the bank's incentives to take risks. Moreover, the existence of guarantees can increase the charter value of a bank, possibly reducing the incentive to take risks as shareholder value is maximized when the bank continues its operations. Cordella and Levy Yeyati (2003) illustrate this problem with a lender of last resort who is able to only pay out in bad states of the world. In practice, it is very difficult to design a "zero moral hazard" policy of bail out or support which is credible ex ante.

There is a plethora of papers which examine this effect, of which a selection are summarized in this paragraph. An influential paper by Keeley (1990) suggests that an increase in competition in the 1980s led bank charter values to decline, which caused banks to take more risk and reduce their capital, increasing their risk of default. But later papers suggest that guarantees can increase risk-taking. Nier and Baumann (2006) examine a panel of banks between 1993 and 2000. They find that, while government safety nets result in lower capital buffers, stronger market discipline resulting from uninsured liabilities and disclosure leads to larger capital buffers, all else being equal. Gropp et al. (2014) exploit a natural experiment to examine the effect of government guarantees on bank risk-taking. Their results suggest that banks whose government guarantees were removed reduced their credit risk by cutting off the riskiest borrowers from credit. They also find that yield spreads of savings banks' bonds increased significantly after the announcement of the decision to remove guarantees, while the yield spread of a sample of bonds issued by a control group remained unchanged. Gropp et al. (2011) use ratings as a proxy for state support and find evidence in favor of the charter value effect. They find no evidence that public guarantees increase the protected banks' risk-taking, but they do find that government guarantees strongly increase the risk-taking of competitor banks.

4 Is market discipline optimal? How much is the right amount?

So far our discussion has focused on whether banks are opaque. This section examines whether transparency or opacity is socially optimal.

In an influential recent paper Dang et al. (2014) examine whether banks are optimally opaque. If a bank's assets are highly transparent, then its market value will fluctuate more often, making its debt liabilities a poorer store of value and thus less useful as a transaction medium. This could be argued as a reason why increasing transparency might not be socially optimal, since money creation is an important social function of banks. However, Gorton and Pennacchi (1990) show that trading losses from information asymmetries can be mitigated by tranching, which should stabilize the value of the most senior liabilities, making them more suitable to use as a transaction medium. They cite bank debt as an example of a type of liquid security which protects relatively uninformed agents, and they provide a rationale for deposit insurance.

The conventional wisdom is that higher transparency via greater disclosure may lead to more market discipline. Goldstein and Sapra (2014) provide an excellent discussion showing analytically that this may not necessarily be the case for banks. This is because banks operate in the "second-best" environment – in other words, the presence of market distortions means that introducing another friction may lead to a more efficient outcome. Examples of such distortions are the interconnected nature of banks, the presence of social externalities, principal-agent problems, and taxes and bankruptcy costs.

Even if greater disclosure boosts market discipline, it may be the case that more effective market discipline does not result in higher economic efficiency. Goldstein and Sapra (2014) show that, although greater disclosure is ex post efficient, this does not necessarily translate into ex ante inefficiencies. They argue that disclosure of stress test information may be beneficial ex post in that it improves market discipline, but if the opacity of the bank's operations means that market participants do not have an adequate understanding of a bank's operations, then market discipline may be hampered by inducing the bank to choose sub-optimal portfolios or inefficient asset sales, thereby reducing economic efficiency. Kleymenova (2013) provides empirical evidence of this in the disclosure of banks' borrowing from the US Federal Reserve's Discount Window during the financial crisis. She finds that these disclosures contained positive information for the market and that they decrease banks' cost of capital. But she also documents how banks change their behavior: banks respond to the discount window disclosures by increasing their liquidity holdings and decreasing risky assets. Thakor (2015) analytically predicts that mandatory disclosure for financial institutions might be inefficient and make banks more fragile. This result comes from banks not wishing to disclose information which may lead to an increase in disagreement. Overall, these two studies argue that, in response to increased mandatory disclosure, banks change their behavior to avoid further disclosures. On the other hand Bischof and Daske (2013) find a substantial and relative increase in stress test participants' voluntary disclosure of sovereign credit risk exposures subsequent to the mandated release of credit risk-related disclosures.

Morris and Shin (2002) show that greater disclosure may be harmful because it induces market participants to put excessive weight on the public information. If the public information is not very precise, then such excessive weight may actually hamper market discipline because market participants rely too much on the non-fundamental or noise component of the disclosure. Similarly, Goldstein and Leitner (2015) examine the issue of opacity, but from the perspective of the regulator. In their setup a regulator has information about banks' ability to overcome future liquidity shocks. Disclosing information may prevent a market breakdown but also destroy risk-sharing opportunities. They show that risk-sharing arrangements work well if the overall state of the financial industry is perceived to be strong. But in bad times, partial disclosure by the regulator can be the optimal solution.

Unfortunately this is easier to pose as a policy than to put into practice. Disclosure standards are slow to change and it is difficult to remove sensitive information from regular reports once investors have become accustomed to it, without creating a perceived signal and increasing further panic. This may explain why strong banks do not necessarily choose to signal their strength by disclosing more, for fear they may become weak banks in future and be unable to stop disclosing this information.

5 Concluding remarks

Insufficient disclosure of information by commercial banks can act as an amplifier of financial stress. It can also distort funding choices and prices in non-crisis times, by contributing to information asymmetry between insiders (such as managers) and outsiders (market participants). Theoretical and empirical evidence suggests that investors do not appear to adequately discipline banks or demand that they disclose the information required to properly assess the risks that they take. This can be partially explained by the "too big to fail" effect, which leads to some of these risks being shared with the government. But there is also evidence to suggest that banks may be too complex for their risks to be properly understood by outside investors.

Mandatory disclosure by regulators can help address some of these asymmetry problems, especially if it helps overcome principal-agent problems, which prevent outside investors from achieving the desired level of disclosure. It is important, however, to consider the effect it may have on the behavior of banks and their investors. Simply publishing as much information as possible is not likely to be helpful to outsiders trying to understand a complex business, and it may actually exacerbate problems during times of stress. Regulation on disclosure can take several forms – for example, information can be published by the firm itself (e.g., enhanced reporting standards) or the regulator (e.g., stress test disclosures) – and these choices may have important consequences for the way that the material is perceived by outsiders at different points in the financial cycle. We do not yet have a good understanding of what policies would optimize social welfare on a time-consistent "through the cycle" basis: this may be an area for further research.

Notes

* Any views expressed are solely those of the authors and so cannot be taken to represent those of the Bank of England or to state Bank of England policy. This paper should therefore not be reported as representing the views of the Bank of England or members of the Monetary Policy Committee or Financial Policy Committee.

1. This may change in future. An exciting new area of research concerns the use of search terms and news articles. For example Drake et al. (2012) examine investors' searches around the days of earning announcements and Drake et al. (2015) examines traffic on the SEC's EDGAR servers.
2. Rating agencies received considerable criticism after the financial crisis for having skewed incentives as the issuer pays the rating agency to rate them, which led them to assign higher ratings to products. This is a valid criticism but should apply to almost all ratings of banks almost equally. This 'uprating' mainly involved structured products for which fees were much larger and for which it was possible to "shop" for ratings. By contrast, for publically traded bank debt, a bank is rated by all rating agencies. See Bolton et al. (2012) for more on the incentives of ratings agencies.
3. An earlier version of these charts – calculated by the authors – can be found in Haldane's speech.
4. One interpretation of this may be that large non-financial firms are also opaque. Cohen and Lou (2012) find evidence that investors take longer to process information for multi-industry conglomerates than for simpler single-industry firms.

References

Acharya, V.V., Anginer, D., and Warburton, A.J. 2014. “The End of Market Discipline?” available at http://pages.stern.nyu.edu/˜sternfin/vacharya/.

Bank of England. 2009. *Financial Stability Report*, December, available at www.bankofengland.co.uk/publications/Documents/fsr/2009/fsrfull0912.pdf.

Bank of England. 2013. “A Framework for Stress Testing the UK Banking System.” *Discussion Paper*, October.

Basel Committee on Banking Supervision. 2006. “International Convergence of Capital Measurement and Capital Standards.” available at http://www.bis.org/publ/bcbs128.pdf.

Bischof, J., and Daske, H. 2013. “Mandatory Disclosure, Voluntary Disclosure, and Stock Market Liquidity: Evidence from the EU Bank Stress Tests.” *Journal of Accounting Research*, 51: 997–1029.

Black, F. and Cox, J.C. 1976. “Valuing Corporate Securities: Some Effects of Bond Indenture Provisions.” *The Journal of Finance*, 31: 351–367.

Board of Governors of the Federal Reserve System and Secretary of the Treasury. 2000. “The Feasibility and Desirability of Mandatory Subordinated Debt.” *Report to Congress*.

Bolton, P., Freixas, X., and Shapiro, J. 2012. “The Credit Ratings Game.” *The Journal of Finance*, 67: 85–111.

Boyd, J.H., and Graham, S.L. 1988. “The Profitability and Risk Effects of Allowing Bank Holding Companies to Merge with Other Financial Firms: A Simulation Study.” *Federal Reserve Bank of Minneapolis Quarterly Review,* 12 (2): 3–20.

Cohen, L., and Lou, D. 2012. “Complicated Firms.” *Journal of Financial Economics*, 104 (2): 383–400.

Cordella, T., and Levy Yeyati, E. 2003. “Bank Bailouts: Moral Hazard vs. Value Effect.” *Journal of Financial Intermediation*, 12 (4): 300–330.

Crockett, A. 2001. “Market Discipline and Financial Stability.” *Bank for International Settlements*, available at www.bis.org/speeches/sp010523.htm.

Dang, T.V., Gorton, G., and Holmström, B. 2013. “The Information Sensitivity of a Security.” *Colombia University Working Paper*.

Dang, T.V., Gorton, G., Holmström, B. and Ordoñez, G. 2014. “Banks as Secret Keepers.” *NBER Working Papers*, 20255.

Diamond, D.W. 1984. “Financial Intermediation and Delegated Monitoring.” *The Review of Financial Studies*, 51 (3): 393–414.

Drake, M.S., Roulstone, D.T., and Thornock, J.R. 2012. “Investor Information Demand: Evidence from Google Searches around Earnings Announcements.” *Journal of Accounting Research*, 50(4): 1001–1040.

Drake, M.S., Roulstone, D.T., and Thornock, J.R. 2015. “The Determinants and Consequences of Information Acquisition via EDGAR.” *Contemporary Accounting Research*, Forthcoming.

Easley, D., and O’Hara, M. 2004. “Information and the Cost of Capital.” *The Journal of Finance*, 59 (4): 1553–1583.

The Economist. 2007. “Black Boxes.” 17th May, retrieved from http://www.economist.com/node/9141547.

Ellahie, A. 2013. “Capital Market Consequences of EU Bank Stress Tests.” *London Business School Working Paper*.

Ellis, D.M., and Flannery, M.J. 1992. “Does the Debt Market Assess Large Banks’ Risk?: Time Series Evidence from Money Center CDs.” *Journal of Monetary Economics*, 30 (3): 481–502.

Flannery, M.J., Kwan, S.H., and Nimalendran, M. 2004. "Market Evidence on the Opaqueness of Banking Firms' Assets." *Journal of Financial Economics*, 71 (3): 419–460.

Flannery, M.J., Kwan, S.H., and Nimalendran, M. 2013. "The 2007–2009 Financial Crisis and Bank Opaqueness." *Journal of Financial Intermediation*, 22 (1): 55–84.

Flannery, M.J., and Sorescu, S.M. 1996. "Evidence of Bank Market Discipline in Subordinated Debenture Yields: 1983–1991." *The Journal of Finance*, 51 (4): 1347–1377.

Glasserman, P., and Tangirala, G. 2015. "Are the Federal Reserve's Stress Test Results Predictable?" *Office of Financial Research working paper*, No. 15–02.

Goldstein, I., and Leitner, Y. 2015. "Stress Tests and Information Disclosure." *Federal Reserve Bank of Philadelphia Working Paper*, No. 15–10.

Goldstein, I., and Sapra, H. 2014. "Should Banks' Stress Test Results be Disclosed? An Analysis of the Costs and Benefits." *Foundations and Trends in Finance*, 8 (1): 1–54.

Gorton, G., and Pennacchi, G. 1990. "Financial Intermediaries and Liquidity Creation." *The Journal of Finance*, 45 (1): 49–71.

Gorton, G., and Santomero, A.M. 1990. "Market Discipline and Bank Subordinated Debt." *Journal of Money, Credit and Banking*, 22 (1): 119–128.

Gropp, R., Hakenes, H., and Schnabel, I. 2011. "Competition, Risk-shifting, and Public Bail-out Policies." *Review of Financial Studies*, 24 (6): 2084–2120.

Gropp, R., Gründl, C., and Guettler, A. 2014. "The Impact of Public Guarantees on Bank Risk Taking: Evidence from a Natural Experiment." *Review of Finance*, 18: 457–488.

Haggard, K.S., and Howe, J.S. 2012. "Are Banks Opaque?" *International Review of Accounting, Banking, and Finance,* 4 (1): 51–72.

Haldane, A.G. 2011. "Capital Discipline." *Speech*.

Iannotta, G. 2006. "Testing for Opaqueness in the European Banking Industry: Evidence from Bond Credit Ratings." *Journal of Financial Services Research*, 30 (3): 287–309.

James, C. 1991. "The Losses Realized in Bank Failures." *The Journal of Finance*, 46: 1223–1242.

Jin, L., and Myers, S.C. (2006), "R-squared around the World: New Theory and New Tests." *Journal of Financial Economics*, 79 (2): 257–292.

Keeley, M. 1990. "Deposit Insurance, Risk, and Market Power in Banking." *American Economic Review*, 80 (5): 1183–1200.

Kleymenova, A. 2013. "Consequences of Mandated Bank Liquidity Disclosures." *SSRN Electronic Journal 01/2013*, DOI: 10.2139/ssrn.2201146.

Kyle, A.S. 1985. "Continuous Auctions and Insider Trading." *Econometrica*, 53 (6): 1315–1335.

Merton, R.C. 1974. "On the Pricing of Corporate Debt: The Risk Structure of Interest Rates." *The Journal of Finance*, 29 (2): 449–470.

Morgan, D. 2002. "Rating Banks: Risk and Uncertainty in an Opaque Industry." *American Economic Review*, 92 (4): 874–888.

Morgan, D.P., Peristiani, S., and Savino, V. 2014. "The Information Value of the Stress Test." *Journal of Money, Credit and Banking*, 46 (7): 1479–1500.

Morgan, D.P., and Stiroh, K.J. 1999. "Bond Market Discipline of Banks: Is the Market Tough Enough?" *Federal Reserve Bank of New York Staff Report*, 95.

Morris, S., and Shin, H.S. 2002. "Social Value of Public Information." *American Economic Review*, 92 (5): 1521–1534.

Myers, S.C. 2000. "Outside Equity." *The Journal of Finance*, 55 (3): 1005–1037.

Myers, S.C., and Majluf, N.S. 1984. "Corporate Financing and Investment Decisions When Firms Have Information that Investors Do Not Have." *Journal of Financial Economics*, 13 (2): 187–221.

Nier, E., and Baumann, U. 2006. "Market Discipline, Disclosure and Moral Hazard in Banking." *Journal of Financial Intermediation*, 15 (3): 332–361.

Siegert, C., and Willison, M. 2015. "Estimating the Extent of the 'Too Big to Fail' Problem – A Review of Existing Approaches." *Bank of England Financial Stability Paper*, 32.

Sironi, A. 2003. "Testing for Market Discipline in the European Banking Industry: Evidence from Subordinated Debt Issues." *Journal of Money, Credit and Banking*, 35 (3): 443–472.

Sowerbutts, R., Zimmerman, P., and Zer, I. 2013. "Banks' Disclosure and Financial Stability." *Bank of England Quarterly Bulletin*, 53 (4): 326–335.

Thakor, A.V. 2015. "Strategic Information Disclosure When There is Fundamental Disagreement." *Journal of Financial Intermediation*, 24 (2): 131–153.

4
Strategic Monetary and Fiscal Policy Interaction in a Liquidity Trap

Ali al-Nowaihi and Sanjit Dhami

1 Introduction

In its classical form, the *liquidity trap*, a term coined by Keynes (1936), is a situation where an increase in money supply fails to reduce the nominal interest rate. The modern literature has concentrated on the case where the nominal interest rate has been driven down to zero (the so called 'zero bound'). The source of a liquidity trap, in most circumstances, is a sharp fall in aggregate demand; see Keynes (1936), Bernanke (2002). Interest in the liquidity trap has revived in recent years due, in no small measure, to the experience of Japan since 1990. Woodford (2005, 29) discusses the near miss of the US economy from a liquidity trap in the summer of 2003. The era of successful delegation of monetary policy to independent central banks with low inflation targets[1] opens up the possibility that sufficiently large negative demand shocks might push an economy into a liquidity trap with huge associated welfare consequences.[2] Blanchard et al. (2010) propose an inflation target of 4 percent in order to provide greater range for the nominal interest rate instrument. Our paper provides one framework within which to evaluate this proposal.

In a liquidity trap traditional monetary policy loses its effectiveness because nominal interest rates can be reduced no further in order to boost the interest sensitive components of aggregate demand. Hence, reliance must be placed on other, possibly more expensive, policies. Keynes (1936), in the first policy prescription for a liquidity trap, suggested the use of fiscal policy, which works through the multiplier effect to boost output and employment.

However, the recent literature has largely focussed on monetary policy and the role of expectations. Krugman (1998, 1999) reformulated the liquidity trap as a situation where an economy requires a negative real interest rate. With nominal interest rates bound below by zero, the only way in which a negative real interest rate can be achieved is to have an expectation of positive inflation.[3]

This, in turn, creates a need for a credible commitment to the future level of actual inflation because after the economy has escaped from the liquidity trap it is in the interest of all parties to reduce inflation. A forward-looking private sector will anticipate this and expect low future inflation. But then the real interest rate remains positive, keeping the economy in a liquidity trap.

The subsequent literature on the liquidity trap has also considered exchange rate policies such as currency depreciation, integral stabilization, a carry tax on currency, open market operations in long term bonds, price level targets, and money growth rate pegs. The surveys in Svensson (2003) and Blinder (2000) consider these policies in detail, however, these policies have important limitations.[4,5]

Eggertsson (2006a, b) recommends abandonment of an independent central bank and a return to discretionary policy by a unitary monetary-fiscal authority. A debt financed fiscal expansion during a liquidity trap results, via the government budget constraint, in higher expectations of future inflation. Eggertsson shows that this solution is superior to either monetary policy alone or uncoordinated monetary and fiscal policy. However, as Eggertsson shows, even optimal discretion is inferior to the fully optimal rational expectations solution with commitment. Moreover, abandoning delegation of monetary policy to an independent central bank with a narrow mandate, in favor of a return to discretion, appears to be a retrograde step.[6]

In this paper, we find that the optimal institutional response to the possibility of a liquidity trap has two main components. First, an optimal inflation target given to the operationally independent Central Bank. Second, the Treasury, who retains control over fiscal policy and acts as leader, is given optimal output and inflation targets. This keeps inflationary expectations sufficiently high and achieves the optimal rational expectations pre-commitment solution. Simulations show that this arrangement is (1) optimal even when the Treasury has no inflation target but follows the optimal output target and (2) 'near optimal' even when the Treasury follows its own agenda through a suboptimal output target but is willing to follow an optimal inflation target. Finally, if monetary policy is delegated to an independent central bank with an optimal inflation target, but the Treasury retains discretion over fiscal policy, then the outcome can be a very poor one.

We also consider a version of our model with a government budget constraint. We find that the optimal solution can be achieved by appropriate inflation and fiscal targets given to the Bank and the Treasury, respectively.

1.1 The Japanese experience: fiscal policy

The Japanese experience with the liquidity trap since the 1990's is now well documented; for instance, Posen (1998). Here we emphasize three points.[7]

J1 *Potency of fiscal policy in a liquidity trap*: The large budget deficits in Japan over the 1990's, with debt reaching a peak of about 140 percent of GDP, have sometimes formed the basis for the conclusion that Japanese fiscal policy was not effective in the liquidity trap. However, this view is at variance with the empirical evidence; for instance Posen (1998), Kuttner and Posen (2001), Iwamura et al. (2005) and Ball (2005). Kuttner and Posen show that tax revenues fell through the deflation of the 1990's. Worried by the special demographic problems faced by Japan, the budget deficits largely funded existing expenditure commitments. It follows that the stabilization component of Japanese fiscal policy in the 1990's was quite weak. Kuttner and Posen show that when the fiscal stimulus was strong, such as in the fiscal package of 1995, it worked in stimulating GDP. On the whole, however, expansionary fiscal policies were largely offset by other contractionary components of fiscal policy such as an increase in the national consumption tax from 3 percent to 5 percent, increase in the contribution rates to social security and the repeal of temporary tax cuts. It is useful to cite more fully from Posen (1998). He writes "The reality of Japanese fiscal policy in the 1990's is less mysterious and ultimately, more disappointing. The actual amount injected into the economy by the Japanese government- through either public spending or tax reductions- was about a third of the total amount announced. This limited quantity of total fiscal stimulus was disbursed in inefficiently sized and inefficiently administered doses with the exception of the 1995 stimulus package. The package did result in solid growth in 1996, demonstrating that fiscal policy does work when it is tried....On net, the Japanese fiscal stance in the 1990's was barely expansionary." The empirical results of Iwamura et al. (2005) and Ball (2005) lend strong support to the finding of Kuttner and Posen. Eggertsson (2006b) calculates a deficit spending multiplier of 3.76, which is much higher than previously thought.

J2 *Lack of appropriate institutions and incentives for policy makers*: The inability of the Japanese Treasury to follow through with an appropriate fiscal stimulus suggests the possibility of inadequate institutional foundations to deal with the liquidity trap. For instance, the Japanese fiscal and monetary authorities did not have any explicit output/inflation targets prior to the onset of the liquidity trap that (1) might have created incentives for an appropriate response, and (2) altered expectations, particularly inflationary expectations, that could have dampened the liquidity trap.

J3 *Lack of coordination between the fiscal and monetary authorities*: Competing policy authorities might disagree on the appropriate response to a liquidity trap, possibly worsening the situation. For instance, the empirical results of Iwamura et al. (2005) indicate lack of coordination between the monetary and fiscal policy authorities. They write "It also suggests that policy coordination between the government and the Bank of Japan did not work well during this period, in the sense that the government deviated from the Ricardian rule towards fiscal tightening while the BOJ (Bank of Japan) adopted a zero interest rate policy and quantitative easing." Eggertsson (2006b) calculates a deficit spending multiplier of exactly zero, for this scenario.

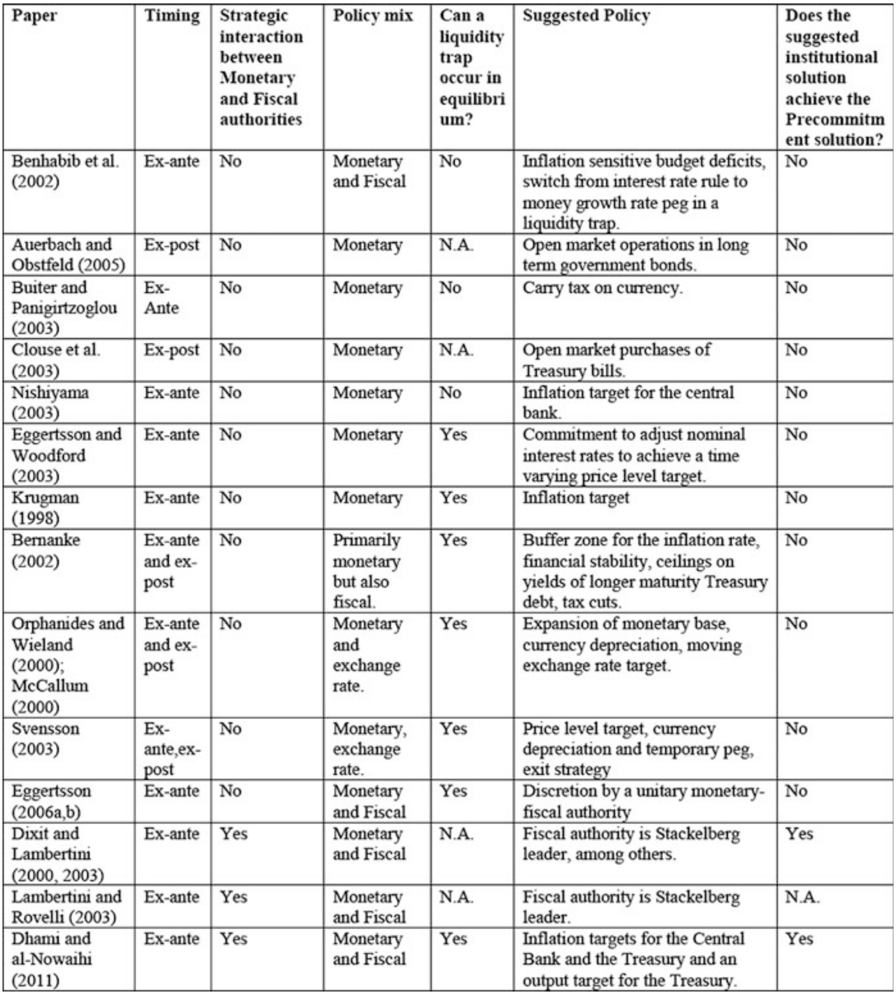
1.2 About our chapter

To motivate our paper we ask the following three questions.

Q1 *Is there strategic policy interaction between the various policy makers?*
Models of strategic monetary and fiscal policy interaction have recently been given a new impetus by the work of Dixit and Lambertini (2003) and Lambertini and Rovelli (2003) (which, however, do not consider a liquidity trap). Issues of strategic interaction between policy makers assume even greater significance during times of extreme recessions as the Japanese experience (J3 above) indicates. However, issues of strategic policy interaction between monetary and fiscal authorities are completely ignored by the theoretical work on the liquidity trap. Typically the only policy considered is monetary policy and so issues of strategic interaction do not arise.[8] On the other hand, when multiple policies are considered, their strategic interaction is not considered.[9]

Q2 *Can liquidity traps occur in equilibrium?*
One strand of the literature considers policies that could mitigate the effects of liquidity traps. The other strand prescribes policies that would prevent the economy from ever falling into a liquidity trap.[10] In general, the optimal policy for our model allows the economy to fall into a liquidity trap with some probability. Thus our model is in the economics tradition that stresses limiting economic bads (e.g., externalities) to their 'optimal level', rather than complete elimination.[11]

Q3 *Is the perspective ex-ante or ex-post?*
The literature typically asks either one of the following two questions. (1) What is the optimal institutional design (assignments of targets and instruments to the various policy makers) when there is the possibility of a liquidity trap in the future? (2) Given that the economy is in a liquidity trap, what actions can be taken to eliminate the liquidity trap.[12] There is considerable disagreement on both questions; particularly the latter. An ex-ante perspective allows one to plan optimally for a problem before it arises, while an ex-post approach is mainly concerned with damage control. Furthermore, the announcements of policy makers during a liquidity trap (an ex-post perspective) might carry little credibility for the public. Hence, ideally one would like to look at the appropriate institutional design prior to the onset of a liquidity trap (an ex-ante perspective).



Paper	Timing	Strategic interaction between Monetary and Fiscal authorities	Policy mix	Can a liquidity trap occur in equilibrium?	Suggested Policy	Does the suggested institutional solution achieve the Precommitment solution?
Benhabib et al. (2002)	Ex-ante	No	Monetary and Fiscal	No	Inflation sensitive budget deficits, switch from interest rate rule to money growth rate peg in a liquidity trap.	No
Auerbach and Obstfeld (2005)	Ex-post	No	Monetary	N.A.	Open market operations in long term government bonds.	No
Buiter and Panigirtzoglou (2003)	Ex-Ante	No	Monetary	No	Carry tax on currency.	No
Clouse et al. (2003)	Ex-post	No	Monetary	N.A.	Open market purchases of Treasury bills.	No
Nishiyama (2003)	Ex-ante	No	Monetary	No	Inflation target for the central bank.	No
Eggertsson and Woodford (2003)	Ex-ante	No	Monetary	Yes	Commitment to adjust nominal interest rates to achieve a time varying price level target.	No
Krugman (1998)	Ex-ante	No	Monetary	Yes	Inflation target	No
Bernanke (2002)	Ex-ante and ex-post	No	Primarily monetary but also fiscal.	Yes	Buffer zone for the inflation rate, financial stability, ceilings on yields of longer maturity Treasury debt, tax cuts.	No
Orphanides and Wieland (2000); McCallum (2000)	Ex-ante and ex-post	No	Monetary and exchange rate.	Yes	Expansion of monetary base, currency depreciation, moving exchange rate target.	No
Svensson (2003)	Ex-ante,ex-post	No	Monetary, exchange rate.	Yes	Price level target, currency depreciation and temporary peg, exit strategy	No
Eggertsson (2006a,b)	Ex-ante	No	Monetary and Fiscal	Yes	Discretion by a unitary monetary-fiscal authority	No
Dixit and Lambertini (2000, 2003)	Ex-ante	Yes	Monetary and Fiscal	N.A.	Fiscal authority is Stackelberg leader, among others.	Yes
Lambertini and Rovelli (2003)	Ex-ante	Yes	Monetary and Fiscal	N.A.	Fiscal authority is Stackelberg leader.	N.A.
Dhami and al-Nowaihi (2011)	Ex-ante	Yes	Monetary and Fiscal	Yes	Inflation targets for the Central Bank and the Treasury and an output target for the Treasury.	Yes

Figure 4.1 Relation of our chapter with the existing literature

We describe our chapter as follows. We would answer yes to the first two questions and 'ex-ante perspective' to the third. We consider strategic interaction between monetary and fiscal authorities in a simple aggregate supply – aggregate demand model similar to the one in Dixit and Lambertini (2003) and Lambertini and Rovelli (2003) but extended to allow for a liquidity trap and the effect of inflationary expectations in the aggregate supply curve. There is some possibility that the economy will fall into a liquidity trap in some state of the world in the future. Our central concern is to identify optimal institutional arrangements[13] from an ex-ante perspective. Figure 4.1 summarizes our paper in relation to the existing literature.

1.3 Some results and intuition

As pointed out above, Krugman identified the solution to a liquidity trap as creating high enough inflationary expectations. However, under discretion, promises of high inflation will not be believed. This is because outside a liquidity trap the correct value for the real interest rate can be achieved more cheaply with zero inflation. Therefore, if the economy turns out not to be liquidity trapped, the Treasury has an incentive to renege on its promise of high inflation. A rational forward looking private sector will anticipate this. The result is low inflation expectations, keeping the real interest rate too high in a liquidity trap. Notice that unlike the standard analysis conducted in the absence of a liquidity trap the discretionary outcome can be suboptimal relative to the precommitment outcome because it creates *too little* inflation (Eggertsson (2006a,b) calls this the deflation bias).

We suggest an institutional solution, the *optimal delegation regime,* that achieves the optimal rational expectations precommitment solution for all parameter values in our model. This regime has three components. First, the Treasury acts as Stackelberg leader and the Central Bank as follower. Second, an inflation target is given to a Central Bank who has exclusive control over monetary policy. Outside a liquidity trap, where monetary policy is effective, the Treasury would rather not use the relatively more costly fiscal stabilization policy, leaving the Central Bank to perform the stabilization function. Because the Central Bank is operationally independent and its sole objective is achieving monetary stability, this type of delegation provides a commitment to the necessary inflation level when the economy is not in a liquidity trap. Our third component is to give the Treasury, who retains control of fiscal policy, something like a *Taylor rule,* which penalizes deviations of output from an output target and inflation from the inflation target. This gives the Treasury the correct incentive to undertake the appropriate (but costly) fiscal stimulus in a liquidity trap where monetary policy is ineffective. Consequently, inflation expectations are at the right level to produce the correct value for the real interest rate in a liquidity trap. For a variety of reasons, e.g., electoral concerns, the output target of the Treasury may differ from the optimal target. In this case, we find that even if the Treasury's output target is substantially different from the optimal output target, this *suboptimal delegation regime* achieves *close* to the optimal solution and is much better than discretion.

While it may appear reasonable to assign an inflation target to the Central Bank, it may be asked why should the Treasury have an inflation target, as well as an output target? To answer this question, we define two further regimes: the *output nutter regime,* where the Treasury has an output target but not an inflation target; and the *reckless output nutter regime* where the Treasury has an output target but does not have an inflation target and does not care about the cost of fiscal policy. It turns out that so long as the Treasury follows the optimal output target, then delegation achieves the optimal solution even in the regimes of the output nutter and the reckless output nutter. However, in the latter two cases, the delegation regime is not robust; in the sense that if the output target of the Treasury is different from the optimal target, then performance is poor and can be much worse than under discretion. Hence, giving the Treasury an inflation target (as well as an output target), while not essential for optimality, adds to the robustness of the policy. In particular the hybrid regime where monetary policy is delegated to an independent central bank with an optimal inflation target, while the Treasury retains discretion over fiscal policy, can perform badly and much worse than had the Treasury retained discretion over both monetary and fiscal policy. We summarize these results in Figure 4.2. In each regime the central bank follows its optimally assigned inflation target.

Furthermore, the *optimal delegation regime* achieves the optimal mix between monetary and fiscal policy as we now explain. Theoretically, society could give a sufficiently high inflation target to the Central Bank which in turn generates sufficiently high inflation expectations so that the nominal interest rate never hits its zero floor. While this policy would always avoid the liquidity trap, it is not optimal because inflation is costly. Analogously it is not optimal to give the Treasury too high an output target because if a liquidity trap occurs, it would use the costly fiscal policy excessively. The optimal solution then is to have a mix of both i.e., some inflation outside a liquidity trap and some dependence on costly fiscal policy in a liquidity trap.

The first best is achieved if one could remove the distortions that cause the liquidity trap. The second best obtains with the optimal rational expectations commitment solution. The third best is achieved with various institutional design features introduced into policy making. The fourth best obtains under discretion. It is well known that, in the absence of a liquidity trap, 'optimal institution design', such as Walsh contracts, can achieve the second best. Our suggested institutional design achieves the second best in the presence of a liquidity trap.

In section 5 we consider a version of our model with a government budget constraint. We find that the optimal solution can be achieved by assigning appropriate inflation targets to the Central Bank and appropriate surplus/deficit targets to the Treasury.

1.4 Optimal control versus game theory

To simplify the dynamic *game-theoretic* analysis we follow the tradition, established in the time-inconsistency literature,[14] of abstracting from *structural* dynamic issues, notably, capital formation, the term structure of interest rates,

Regime	Treasury follows optimal Output target	Treasury follows optimal inflation target	Treasury cares about inflation	Treasury follows personal output agenda	Outcome
Optimal Delegation	Yes	Yes	Yes	No	Precommitment Solution
Suboptimal Delegation	No	Yes	Yes	Yes	Near optimal
Output nutter	Yes	No	No	No	Precommitment Solution
Reckless nutter	Yes	No	No	No	Precommitment Solution
Output nutter	No	No	Yes	Yes	Much worse than discretion
Reckless nutter	No	No	No	Yes	Much worse than discretion

Figure 4.2 Outcomes under various regimes

exchange rate policy and the financing of the stabilization component of fiscal policy. Concentrating on the aggregate demand consequences of investment expenditure, but abstracting from its contribution to growth, is standard in models of the business cycle, and is a feature of all the models of the liquidity trap (as far as we know).

Eggertsson and Woodford (2003), in a structurally dynamic model of monetary policy with a financial sector and a zero lower bound on interest rates, show that the short-run interest rate (which is the instrument of policy) determines all other interest rates and exchange rates. As they clearly explain, open market operations only work to the extent that they enhance the credibility of policy. Thus, and in common with many models, we take the short-run interest rate as directly affecting aggregate demand and we abstract from open economy aspects.

Except for section 5, we do not explicitly model the government budget constraint. This does not necessarily imply that the government budget constraint is violated. For the government could run a fiscal surplus outside a liquidity trap. This could then finance a fiscal deficit in a liquidity trap. Section 5 explicitly models the government budget constraint and shows that the qualitative results of our paper are not changed.

Nevertheless, we incorporate an element of structural dynamics resulting from persistence in demand shocks (Section 6). We believe that our model thus reproduces the essentials of the problems associated with a liquidity trap: persistence, credibility and monetary-fiscal coordination, in a clear and simple way.

1.5 Relation to Dhami and al-Nowaihi (2011)

Dhami and al-Nowaihi (2011), henceforth DaN, also propose a model of a liquidity trap along the lines that are mentioned above. In this paper we extend their model along the following lines.

E1. Here, we introduce the full set of parameter values. By contrast, DaN assign the value one to all parameters. The advantage of that special choice of parameter values is that all the details of all the proofs can be exhibited. Unfortunately, this is no more the case when the full set of parameter values is introduced, as is done here. While we can still explicitly state the assumptions and the conclusions, the details of the proofs can no longer be printed. The reason is that many of the algebraic expressions are more than one page in length each! However, the logic of the proofs here is the same as in DaN. All proofs require only elementary (though tedious) algebraic calculations. All our claims can be independently checked by a reader wishing to reconstruct the intermediate steps of the calculations or willing to use the ‘check equality’ command of a scientific word processor.

E2. Here, we allow for persistence in demand shocks. By contrast, in DaN the demand shocks are uncorrelated over time.

E3. Here, we allow for general probability distribution over the two states of the world. By contrast, in DaN the demand shocks in any period take two possible values, 1 and −1 with equal probability.

E4. Here, we show that if the Treasury follows its own private output target, y_T, rather than the optimal output target, y_T^*, then the resulting 'suboptimal delegation regime' is, nevertheless, close to the 'optimal delegation regime' and is much better than discretion. This analysis is absent in DaN.

E5. Here, we show that giving the Treasury an inflation target (as well as an output target), while not essential for optimality, adds to the robustness of the policy. In particular the hybrid regime where monetary policy is delegated to an independent central bank with an optimal inflation target, while the Treasury retains discretion over fiscal policy, can perform badly and much worse than had the Treasury retained *discretion* over both monetary and fiscal policy. This analysis is absent in DaN.

E6. In section 5 we consider a version of our model with a government budget constraint. This is absent in DaN.

1.6 Schematic outline

The model is formulated in Section 2. Section 3 derives the two benchmark solutions: the *optimal rational expectations precommitment solution* and the *discretionary solution*. Section 4 derives the *optimal delegation solution*. Section 5 considers a version of our model with a government budget constraint. Section 6 demonstrates the robustness of the model by allowing for the full set of parameters, persistence of demand shocks and several alternative formulations of the Treasury's objectives. Section 7 discusses the relation of our paper to the literature. Section 8 concludes with a brief summary. Proofs are relegated to appendices.

2 Model

In this section we describe the most parsimonious version of the model. In Section 6 below, we demonstrate the robustness of the results of this model with respect to the full set of parameters, persistent demand shocks, a general probability distribution over the two states of nature, and further considerations about the Treasury's objectives.

2.1 Aggregate demand and aggregate supply

We use an aggregate demand and supply framework that is similar to Ball (2005), Dixit and Lambertini (2003) and Lambertini and Rovelli (2003). The aggregate demand and supply equations are given by, respectively

$$AD : y = f - \left(i - \pi^e\right) + \epsilon \tag{4.1}$$

$$AS : y = \pi - \pi^e \tag{4.2}$$

where y is the deviation of output from the natural rate and f captures fiscal policy.[15] For example, $f > 0$ could denote a fiscal deficit (either debt financed or money financed[16]) while $f < 0$, a fiscal surplus. But f could also denote a temporary balanced budget reallocation of taxes and subsidies that has a net expansionary effect; for instance Dixit and Lambertini (2000). $i \geq 0$ is the nominal interest rate, π is the rate of inflation, π^e is expected inflation[17] and ϵ is a demand shock.[18] The instruments of policy are i and f. The demand shock ϵ takes two values, $a, -a$, with equal probability, where $a > 0$, hence

$$E[\epsilon] = 0,\ Var[\epsilon] = a^2. \tag{4.3}$$

The aggregate demand equation reflects the fact that demand is increasing in the fiscal impulse, f, and decreasing in the real interest rate; it is also affected by demand shocks. The aggregate supply equation shows that deviations of output from the natural rate are caused by unexpected movements in the rate of inflation. Note the absence of parameters in (4.1), (4.2). This is because our conclusions do not qualitatively depend on the values of such parameters (see Section 6). So we have suppressed them to improve readability.

Equating aggregate demand and supply we get from (4.1) and (4.2), our reduced form equations for output and inflation.

$$y = f - i + \pi^e + \epsilon \tag{4.4}$$

$$\pi = f - i + 2\pi^e + \epsilon \tag{4.5}$$

Hence, fiscal policy, monetary policy and inflation expectations (in the spirit of New Keynesian models) have an affect on output (and so also on unemployment) and inflation.

2.2 Microfoundations

Our model is inspired by the microfounded dynamic model of monopolistic competition and staggered price setting in Dixit and Lambertini (2000, 2003). Our structural model in (4.1), (4.2) (or its variant with the full set of parameters given in (4.17), (4.18) below) is similar to Dixit and Lambertini.[19] In the Dixit and Lambertini framework, unexpected movements in inflation have real effects because prices are staggered. Alternatively, a range of 'rational inattention' theories currently compete as potential explanations for the presence of the unexpected inflation term in (4.2). For instance, see Sims (2003).[20]

2.3 Notation

We shall write a variable with a subscript (sometimes a superscript) '+', for example, y_+, to denote the realization of that variable in the (good) state of the world, $\epsilon = a$. Analogously, to denote the realization of the same variable in the (bad) state of the world, $\epsilon = -a$, we use a subscript (sometimes a superscript) '−', for example, y_-.

2.4 Social preferences

Society's preferences over output and inflation are given by the social welfare function,

$$U_S = -\frac{1}{2}(y - y_S)^2 - \frac{1}{2}\pi^2 - f^2. \tag{4.6}$$

The first term shows that departures of output from its desired level, y_S (note that y_S is the difference between desired output and the natural rate), are costly. We assume that

$$y_S \geq 0 \tag{4.7}$$

This captures the fact that the natural level of output is socially suboptimal (unless $y_S = 0$).[21]

The second term in (4.6) indicates that inflation reduces social welfare. The third term captures the fact that the exercise of fiscal policy is more costly than that of monetary policy.[22] We model this as imposing a strictly positive cost of fiscal policy, f^2, but no cost of using the monetary policy.[23] The cost of using fiscal policy could include deadweight losses, as in Dixit and Lambertini (2003), costs of servicing debt and a risk premium for default.

From (4.6) we see that the first best obtains when $\pi = 0$, $f = 0$, and $y = y_S$. However, from (4.1) and (4.2), it follows that this cannot be an outcome of a rational expectations equilibrium (unless $y_S = 0$).

For expositional clarity we omit parameters in (4.6), but see Section 6. On the microfoundations of such a social welfare function, see Dixit and Lambertini (2000, 2003), Rotemberg and Woodford (1999).

2.4.1 Treasury and social preferences

We will assume for now that society can, if it desires, delegate policy to a "Treasury" that fully internalizes its objective function given in (4.6). So we will use society and Treasury interchangeably here. Other assumptions are considered in Section 6 below.

2.5 Sequence of moves

At the first stage the economy designs its institutions, which assign powers of policy-making decisions to one or two independent policy makers. This is followed by the formation of inflationary expectations, π^e, and the signing of nominal wage contracts in anticipation of future inflation. Next, the demand shock, ϵ, is realized. In light of the actual realization of the shock, the relevant policy makers then decide on the optimal values of the policy variables, f and i. We shall also derive the optimal rational expectations solution (precommitment benchmark) in which the last stage is conducted up-front i.e. the (state contingent) policy variables f and i are announced to the economy prior to the resolution of demand uncertainty.

3 The precommitment and discretionary solutions

3.1 The precommitment regime (The optimal rational expectations solution)

In this section we calculate the globally optimal solution in the class of all rational expectations solutions.[24] The global optimality of the precommitment solution serves as a useful benchmark. The sequence of moves is described below.

The solution method is to find state contingent rules for the policy variables, $i(\epsilon)$, $f(\epsilon)$, i.e., (i_-,f_-), (i_+,f_+), that maximize the expected value of the social welfare (4.6) under the constraints (4.4), (4.5) and the rational expectations condition $\pi^e = E[\pi]$, i.e.

$$\pi^e = \frac{1}{2}\pi_- + \frac{1}{2}\pi_+ \tag{4.8}$$

The results are summarized in Proposition 1. Superscript 'e' denotes expected value.

Proposition 1: *The optimal state-contingent rational expectations precommitment solution is given by*

$\boldsymbol{\epsilon_- = -a < 0}$	$\boldsymbol{\epsilon_+ = a > 0}$	$\boldsymbol{\epsilon^e = 0}$
$i_- = 0$	$i_+ = \frac{6}{5}a$	$i^e = \frac{3}{5}a$
$f_- = \frac{2}{5}a$	$f_+ = 0$	$f^e = \frac{1}{5}a$
$y_- = -\frac{1}{5}a$	$y_+ = \frac{1}{5}a$	$y^e = 0$
$\pi_- = \frac{1}{5}a$	$\pi_+ = \frac{3}{5}a$	$\pi^e = \frac{2}{5}a$
$i_- - \pi^e = -\frac{2}{5}a$	$i_+ - \pi^e = \frac{4}{5}a$	$i^e - \pi^e = \frac{1}{5}a$

The expected utility in the precommitment regime is given by $E\left[U_S^{Opt}\right] = -\frac{1}{5}a^2 - \frac{1}{2}y_S^2$. *Furthermore,* $\left(\frac{\partial U_S}{\partial i}\right)_{Opt} < 0$ *when* $\epsilon = -a$ *and* $\left(\frac{\partial U_S}{\partial i}\right)_{Opt} = 0$ *when* $\epsilon = a$. ■

From Proposition 1 note that $\left(\frac{\partial U_S}{\partial i}\right)_{Opt} < 0$ when $\epsilon = -a$. Hence, the economy is always liquidity trapped when $\epsilon = -a$. In this case, monetary policy is not effective, $i_- = 0$. Hence, the government must commit to using expensive fiscal policy, $f_- = \frac{2}{5}a$, in order to 'lean against the wind'. By contrast, when $\epsilon = a$, monetary policy is effective, $i_+ = \frac{6}{5}a$, and the government has no need for the expensive fiscal instrument, $f_+ = 0$.[25]

Also note that output is below the natural rate (which is normalized to zero) in the liquidity trap ($\epsilon = -a$) but above it otherwise ($\epsilon = a$). On average, it equals the natural rate (recall that y measures the deviation of output from the natural rate). Inflation is positive in both states of the world. The real interest rate is negative[26] in the liquidity trap but positive otherwise and on average.

Recalling that $Var[\epsilon] = a^2$, on average, ceteris paribus, inflation, interest rates and the fiscal instrument of the government will display greater variability in economies where demand shocks have a greater variance and precommitment is possible. Furthermore, the magnitude of policy instruments employed in the two states of the world, $f_- = \frac{2}{5}a$ and $i_+ = \frac{6}{5}a$, are increasing in the size of the shock. This is not surprising as each of these policies fulfills a stabilization role and a larger shock elicits a greater effort in "leaning against the wind".

The solution is independent of y_S, society's desired output relative to the natural rate. As in time consistency models in the absence of the liquidity trap, this occurs because, even if society has a high y_S, the precommitment technology allows it to counter expectations of ex-post surprise inflation (designed to push output towards the high target).

The magnitude of social welfare in this regime depends negatively on the variance of shocks hitting the economy, a^2, and also on the output target of society, y_S.

Finally, note that the values of i_+, i_-, f_+, f_- of the instruments are optimal *ex-ante*. However, after the realization of the shock, $\epsilon = -a$ or $\epsilon = a$, the *ex-post* optimal values of i, f will, in general, be different from these. Thus, for successful implementation, this optimal rational expectations solution needs a precommitment technology. We discuss this in Section 4 below. Next we turn to the second regime in the paper: Discretion.

3.2 Discretionary regime

In this case, the monetary instrument, i, and the fiscal instrument, f, are both assigned to the Treasury. We calculate the time consistent discretionary policy. The sequence of moves is described below.

To find the discretionary solution, first find state-contingent values of the policy variables $i_-(\pi^e), f_-(\pi^e)$ and $i_+(\pi^e), f_+(\pi^e)$ that maximize social welfare (4.6) under the constraints (4.1), (4.2) and conditional on given π^e, ϵ. This allows the computation of the state-contingent inflation rates $\pi_-(\pi^e)$ and $\pi_+(\pi^e)$. Then one needs to find the fixed-point π^e by solving $\pi^e = E[\pi]$:

$$\pi^e = \frac{1}{2}\pi_-\left(\pi^e\right) + \frac{1}{2}\pi_+\left(\pi^e\right) \tag{4.9}$$

Finally, substitute the value for π^e back into the state-contingent policy variables $i_-(\pi^e), f_-(\pi^e)$ and $i_+(\pi^e), f_+(\pi^e)$ to find the solution under discretion.

Depending on the parameter values, a liquidity trap may or may not arise. Proposition 2 below summarizes the results when a liquidity trap, which is the focus of this paper, arises.[27]

Proposition 2: *For $\frac{1}{2}a \leq y_S < a$, the economy is liquidity trapped for $\epsilon = -a < 0$ but not liquidity trapped for $\epsilon = a > 0$. The solution under discretion is given by*

$\boldsymbol{\epsilon_- = -a < 0}$	$\boldsymbol{\epsilon_+ = a > 0}$	$\boldsymbol{\epsilon^e = 0}$
$i_- = 0$	$i_+ = 4y_S - 2a$	$i^e = 2y_S - a$
$f_- = 2(a - y_S) > 0$	$f_+ = 0$	$f^e = (a - y_S) > 0$
$y_- = y_S - a < 0$	$y_+ = a - y_S > 0$	$y^e = 0$
$\pi_- = 4y_S - 3a$	$\pi_+ = 2y_S - a$	$\pi^e = 3y_S - 2a$
$i_- - \pi^e = 2a - 3y_S$	$i_+ - \pi^e = y_S > 0$	$i^e - \pi^e = a - y_S > 0$

and the expected social welfare is given by $E\left[U_S^{Disc}\right] = 12ay_S - 8y_S^2 - 5a^2$

For stabilization purposes, the costly fiscal policy is used only in a liquidity trap when the monetary policy looses effectiveness. As in the precommitment solution, deviations of output from the natural rate are zero on average i.e. $y^e = 0$. The following corollary compares expected social welfare under *Precommitment* with that under *Discretion.*

Corollary 1: *For* $\frac{1}{2}a \leq y_S < a$, $E\left[U_S^{Opt}\right] - E\left[U_S^{Disc}\right] = \frac{3}{10}(5y_S - 4a)^2 \geq 0.$

As one would expect, the presence of a liquidity trap does not alter the ranking between the Precommitment and the Discretion regimes, from a social welfare point of view.

3.3 Alice through the looking glass

Krugman (1998) observed that 'applying conventional modelling to liquidity trap conditions produces unconventional conclusions and policy recommendations'. To which he added (1999) 'The whole subject of the liquidity trap has a sort of Alice-through-the-looking-glass quality'. And indeed, our model exhibits these features, as we will now see.

3.3.1 Precommitment can have higher inflation than discretionary

In the traditional time inconsistency literature, in the absence of a liquidity trap, the optimal level of average inflation is zero (given the welfare function (4.6)) while under discretion it is positive (unless $y_S = 0$, in which case it is also zero); as is well known. The reason is that under discretion, agents perceive (correctly) that the government has an ex-post incentive to create surprise inflation, while under precommitment ex-post surprise inflation is institutionally ruled out.

When a liquidity trap occurs with a positive probability this changes dramatically. From Proposition 1 we see that the optimal level of average inflation under precommitment now is positive ($\pi^e = \frac{2a}{5}$), rather than zero. Under discretion π^e depends on y_S. For $y_S = \frac{1}{2}a$, Proposition 2 gives a negative average expected inflation rate ($\pi^e = -\frac{1}{2}a$), rather than a positive one. Eggertsson (2006a, b) calls this the deflation bias.

The intuitive explanation is as follows. Under precommitment, it is optimal to have positive inflation on average ($\pi^e = \frac{2a}{5}$), despite its cost, to be able to deliver negative real interest rates ($i_- - \pi^e = -\frac{2a}{5}$) in the bad state of the world ($\epsilon = -a$). However, this optimal policy is time inconsistent. If ex-post, the economy is in the good state ($\epsilon = a$) then the optimal real interest rate is positive ($i_+ - \pi^e = \frac{4a}{5}$) which can be achieved more cheaply with zero inflation. Hence, the policy maker has the incentive to renege on its commitment to positive inflation. The rational private sector will perceive this and expect low future inflation. This destroys the credibility of the announcement of high inflation, unless a commitment technology is available.

3.3.2 Higher output targets are a good thing

In the standard textbook model in the absence of a liquidity trap, a higher value of desired output relative to the natural rate, $y_S > 0$, is bad because it leads to high inflation and no gain in output ($y^e = 0$). The reverse occurs with a liquidity trap, $y_S > 0$ is now good! The intuition is that a higher y_S increases inflationary expectations (see Proposition 2) which, by reducing the real interest rate in a liquidity trap, reduces the need for using the expensive fiscal instrument.

If society has a high enough output target (and the Treasury follows it) then, in the discretionary regime, ex-post, a liquidity trap will not arise. However, this outcome might require using the costly fiscal instrument excessively, which could be suboptimal. In Section 4, below, we show this to be precisely the case.

4 Institutions and delegation

In the delegation regime considered in this section, society gives the Central Bank the mandate of achieving an inflation target π_B. The monetary instrument, which is the nominal interest rate, i, is assigned to the Central Bank whose objective is to attain the inflation target π_B. We formalize this by assigning the following objective function to the Central Bank:

$$U_B = -\frac{1}{2}(\pi - \pi_B)^2 \tag{4.10}$$

The fiscal instrument, f, is controlled by the Treasury whose objective function is similar to that of society (4.6) but with, possibly, different inflation and output targets:

$$U_T = -\frac{1}{2}(y - y_T)^2 - \frac{1}{2}(\pi - \pi_T)^2 - f^2 \qquad (4.11)$$

where y_T, π_T are the output and inflation targets respectively of the Treasury. It is important to bear in mind the difference between the socially desirable output level, y_S, and the Treasury's output target, y_T. The optimal value, y_T^*, of y_T, i.e., the value of y_T that maximizes expected social welfare, might be very different from y_S. In fact, our simulations show that y_T^* is well below y_S. Thus a fiscal authority should be 'conservative' in the sense that it should aim for a lower output target than that desired by society, as in Rogoff (1985). See, for example, Table 4.1, below.

4.1 The optimal delegation regime

Under optimal delegation, the game has five stages, shown in Figure 4.3.

The Treasury acts as Stackelberg leader with an output target, y_T, and an inflation target π_T. The Central Bank is the follower with an inflation target π_B. In this subsection we consider the case $\pi_T = \pi_B$ (section 6, below, allows $\pi_T \neq \pi_B$). The Central Bank sets monetary policy taking the fiscal policy, set by the Treasury, as given. The Treasury sets fiscal policy, taking into account the anticipated response of the Central Bank. We solve the game backwards. First we obtain the Central Bank's reaction function $i = i(\pi_B, \pi^e, f, \epsilon)$ by maximizing U_B. Second, we find the Treasury's reaction function $f = f\left(y_T,\ \pi_B, \pi^e,\ \epsilon\right)$ by maximizing U_T. This allows us to derive output and inflation as functions of y_T, $\pi_{B,}$ π^e, ϵ. Third, we determine π^e, assuming rational expectations on the part of the private sector. Fourth, we find the expected social welfare, EU_S, as a function of y_T, π_B, which we maximize to find the optimal values of y_T, π_B which are denoted by y_T^*, π_B^*. We assume that the Treasury and Central Bank adopt the optimal inflation target, π_B^*, and that the Treasury fully complies with the optimal output target, y_T^*. Section 6, below, explores the possibility that the Treasury might not care for inflation and/ or be unwilling to follow the optimal output target, y_T^*, because it has its own output target, y_T. For ease of reference, these concepts are summarized in the following definition.

Definition 1: *y_S is the output level preferred by society (0 is the inflation level preferred by society, see (4.6)). y_T and π_T are output and inflation targets for the Treasury. π_B is the inflation target for the Central Bank. y_T^* and π_B^* are the values of y_T and π_B that maximize expected social welfare, EU_S, subject to the constraints of the model, where U_S is given by (4.6). In section 6, below, we allow the Treasury to adopt an output target, y_T, different from y_T^*, consistent with its own agenda.*

Proposition 3: *Assume that monetary policy is delegated to an independent central bank with inflation target $\pi_B^* = \frac{3}{5}a$. Fiscal policy is retained by the Treasury with output target $y_T^* = \frac{1}{5}a$ and acts as Stackelberg leader. Then the optimal rational* *expectations (precommitment) solution (see Proposition 1) is achieved. Society's expected utility in the optimal delegation regime is given by* $E\left[U_S^{OD}\right] = -\frac{1}{5}a^2 - \frac{1}{2}y_S^2$. *The economy is liquidity trapped only under adverse demand shocks. Inflation and output targets are achieved in the good state but not in the bad state.*[28]

So why does the optimal delegation regime perform so well? The inflation target given to the Central Bank provides a commitment to the necessary inflation level when the economy is not in a liquidity trap. This affects the (ex-ante) inflation expectations which also apply to the (ex-post) liquidity trap ensuring the correct value for the real interest rate in a liquidity trap. Furthermore, inflationary expectations are also influenced correctly by the output and inflation targets given to the Treasury that provide it with the incentive to use the appropriate level of fiscal policy in a liquidity trap. Such an institutional regime achieves the optimal balance between fiscal and monetary policy by neither having to rely too much on costly inflation outside the liquidity trap nor relying too much on costly fiscal policy in a liquidity trap.

5 The government budget constraint

Recall from Propositions 1 and 2 that the optimal solution, and hence optimal delegation, specify a fiscal deficit in a liquidity trap but a fiscal balance outside of a liquidity trap (also recall footnote 15). Here we explicitly model the government budget constraint. We require the Government run a fiscal surplus outside the liquidity trap in order to finance the fiscal deficit in a liquidity trap; so that the government budget constraint holds ex-ante or on average. To facilitate this, to Equations (4.4), (4.5) and (4.8) we here add the government budget constraint:

$$f_- = f, f_+ = -f; \text{ hence, } f_- + f_+ = 0. \tag{4.12}$$

In Proposition 4, below, we derive the optimal state-contingent rational expectations precommitment solution under the government budget constraint (4.12). This is the analog of Proposition 1. The reader may wish to refer back to Figure 4.1, which gives the sequence of moves for the precommitment regime. Subsection 5.1 then describes the optimal delegation regime that

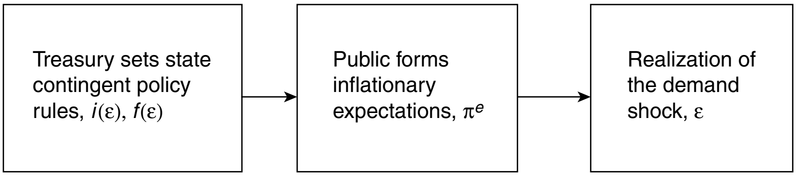

Treasury sets state contingent policy rules, $i(\varepsilon)$, $f(\varepsilon)$
Public forms inflationary expectations, π^e
Realization of the demand shock, ε

Figure 4.3 Sequence of moves for the precommitment regime

implements this precommitment solution. That the optimal delegation indeed implements the precommitment solution is established by Proposition 5.

Proposition 4: *The optimal state-contingent rational expectations precommitment solution under the government budget constraint (4.12) is given by*

$\epsilon_- = -a < 0$	$\epsilon_+ = a > 0$	$\epsilon^e = 0$
$i_- = 0$	$i_+ = a$	$i^e = \frac{1}{2}a$
$f_- = \frac{1}{4}a$	$f_+ = -\frac{1}{4}a$	$f^e = 0$
$y_- = -\frac{1}{4}a$	$y_+ = \frac{1}{4}a$	$y^e = 0$
$\pi_- = \frac{1}{4}a$	$\pi_+ = \frac{3}{4}a$	$\pi^e = \frac{1}{2}a$
$i_- - \pi^e = -\frac{1}{2}a$	$i_+ - \pi^e = \frac{1}{2}a$	$i^e - \pi^e = 0$

The expected utility in the precommitment regime is given by $E\left[U_S^{Opt}\right] = -\frac{1}{4}a^2 - \frac{1}{2}y_S^2$. *Furthermore,* $\left(\frac{\partial U_S}{\partial i}\right)_{Opt} < 0$ *when* $\epsilon = -a$ *and* $\left(\frac{\partial U_S}{\partial i}\right)_{Opt} = 0$ *when* $\epsilon = a$. ■

From Proposition 4 note that $\left(\frac{\partial U_S}{\partial i}\right)_{Opt} < 0$ when $\epsilon = -a$. Hence, the economy is always liquidity trapped when $\epsilon = -a$. In this case, monetary policy is not effective, $i_- = 0$. Hence, the government must commit to using expensive fiscal policy, $f_- = \frac{1}{4}a$, in order to 'lean against the wind'. By contrast, when $\epsilon = a$, monetary policy is effective, $i_+ = a$. However, here, and unlike the case of Proposition 1, the government still needs to run a budget surplus, $f_+ = -\frac{1}{4}a$, to pay for the budget deficit in a liquidity tap. This is, of course, a consequence of the government budget constraint (4.12).

Note that here discretion fails to implement the optimal solution of Proposition 4 for *two* reasons. First, outside the liquidity trap, the government has an incentive to renege on the inflation target (recall subsection 3.3.1). Second, outside a liquidity trap the government has an incentive to renege on its commitment to run a budget surplus.

5.1 The optimal delegation regime

Society assigns to the Treasury the fiscal targets f_T^+, f_T^- to be achieved in the good state, $\epsilon_+ = a$, and the bad state, $\epsilon_- = -a$, respectively. Thus we give the Treasury the objective function

$$U_T^+ = -\frac{1}{2}\left(f_+ - f_T^+\right)^2, \quad \text{if } \epsilon = a, \tag{4.13}$$

$$U_T^- = -\frac{1}{2}\left(f_- - f_T^-\right)^2, \quad \text{if } \epsilon = -a. \tag{4.14}$$

Likewise, society assigns to the Bank the inflation targets π_B^+, π_B^- to be achieved in the good state, $\epsilon_+ = a$, and the bad state, $\epsilon_- = -a$, respectively.

Thus we give the Bank the objective function

$$U_B^+ = -\frac{1}{2}\left(\pi_+ - \pi_B^+\right)^2, \quad \text{if } \epsilon = a, \tag{4.15}$$

$$U_B^- = -\frac{1}{2}\left(\pi_- - \pi_B^-\right)^2, \quad \text{if } \epsilon = -a. \tag{4.16}$$

Under optimal delegation, the game has five stages, shown in Figure 4.6.

The Treasury acts as Stackelberg leader with a surplus target of f_T^+ to be achieved in the good state, $\epsilon = a$, and a deficit target of f_T^- to be achieved in the bad state, $\epsilon = -a$. These targets are chosen partly to guarantee that the government budget constraint holds ex-ante or on average. The Central Bank is the follower with an inflation target π_B^+ to be achieved in the good state, $\epsilon = a$, and an inflation target π_B^- to be achieved in the bad state, $\epsilon = -a$. Announcing these inflation targets helps the public form its rational expectations of the future level of inflation, $\pi^e\left(\pi_B^+, \pi_B^-\right)$. The shock $\epsilon = \pm a$ is then realized. The Central Bank sets monetary policy (i.e., sets interest rates) taking the fiscal policy (i.e., deficits/surpluses), set by the Treasury, as given. The Treasury sets fiscal policy (deficit or surplus). We solve the game backwards. First we obtain the Central Bank's reaction function $i = i\left(\pi_B^+, \pi_B^-, \pi^e, f_+, f_-, \epsilon\right)$ by maximizing its objective function, $U_B^\pm$. Second, we find the Treasury's reaction function $f = f\left(f_T^+, f_T^-, \epsilon\right)$ by maximizing its objective function, $U_T^\pm$. Given these reaction functions, the behavioral Equations (4.4), (4.5) and (4.8), the government budget constraint (4.12) and the stochastic process of the shocks, ϵ, this allows us to find the expected social welfare, EU_S, where U_S is given by (4.6), as a function of the targets $\pi_B^+, \pi_B^-, \pi^e, f_T^+, f_T^-$. Finally, we can find the values of these targets that maximize expected social welfare, EU_S.

Proposition 5: *Assume that monetary policy is delegated to an independent central bank with inflation target $\pi_B^+ = \frac{3}{4}a$, $\pi_B^- = \frac{1}{4}a$, which acts as Stackelberg follower. Fiscal policy is retained by the Treasury with fiscal target $f_T^+ = -\frac{1}{4}a$, $f_T^- = \frac{1}{4}a$ and acts as Stackelberg leader. Then the optimal rational expectations (precommitment) solution (see Proposition 4) is achieved:*

$\boldsymbol{\epsilon_- = -a < 0}$	$\boldsymbol{\epsilon_+ = a > 0}$	$\boldsymbol{\epsilon^e = 0}$
$i_- = 0$	$i_+ = a$	$i^e = \frac{1}{2}a$
$f_- = \frac{1}{4}a$	$f_+ = -\frac{1}{4}a$	$f^e = 0$
$y_- = -\frac{1}{4}a$	$y_+ = \frac{1}{4}a$	$y^e = 0$
$\pi_- = \frac{1}{4}a$	$\pi_+ = \frac{3}{4}a$	$\pi^e = \frac{1}{2}a$
$i_- - \pi^e = -\frac{1}{2}a$	$i_+ - \pi^e = \frac{1}{2}a$	$i^e - \pi^e = 0$

Society's expected utility in the optimal delegation regime is given by $E\left[U_S^{OD}\right] = -\frac{1}{4}a^2 - \frac{1}{2}y_S^2$. The economy is liquidity trapped only under the adverse shock $\epsilon_- = -a$. Inflation and fiscal targets are achieved in the good state and in the bad state.

So why does the optimal delegation regime perform so well? The inflation targets given to the Central Bank provide a commitment to the necessary inflation levels when the economy is in or out of a liquidity trap. This also pins down the (ex-ante) inflation expectations. Together with the fiscal targets, these ensure the correct values for outputs and interest rates in and out of a liquidity trap.

6 The general model

How are our results altered when we introduce the full set of parameters in the model of sections 2–4 and also allow for persistence in the demand shocks with a general probability distribution? What if the Treasury has its own agenda, perhaps on account of electoral concerns or other political economy considerations such as lobbying or interest groups? These issues are considered in this section. We demonstrate that the results of our model are robust to the following five extensions.

E1. Introduction of the full set of parameters.
E2. Persistent demand shocks.
E3. General probability distribution over the two states of the world.
E4. The Treasury might follow an output target, y_T, different from the optimal output target, y_T^*. Recall that y_T^* will, in general, be different from the output level, y_S, most desired by society.
E5. The Treasury and the Central Bank can have distinct inflation targets i.e. $\pi_T \neq \pi_B$.

6.1 A note on output and inflation targets

6.1.1 Inflation targets

There are two main cases. The inflation targets of the Treasury and Central Bank either coincide (i.e. $\pi_T = \pi_B$), or differ (i.e. $\pi_T \neq \pi_B$). In Section 4 we restricted attention to the case $\pi_T = \pi_B$. However, in Subsection 6.5, both cases i.e. $\pi_T = \pi_B$ and $\pi_T \neq \pi_B$ are considered. We show that the optimal delegation regime works equally well in each of these two cases and achieves the optimal rational expectations precommitment solution. Whilst this does not have implications for the optimality of our suggested delegation regime we find the case $\pi_T = \pi_B$ more natural and easier to interpret. Furthermore, we show in Subsection 6.8 that the optimal rational expectations solution can also be achieved if the central bank *alone* has an inflation target while the Treasury simply follows the optimal output target given to it by society.

6.1.2 Output targets

The Treasury is an arm of the government. If the natural rate of output is socially suboptimal, say on account of monopolistic competition, then the government may have an incentive to use fiscal instruments to increase output beyond its natural rate, at least temporarily and a rational private sector will foresee this. The problem of assigning output targets is compounded by the difficulty of measuring deviation of output from its natural rate (compared with the lesser difficulty of measuring deviation of inflation from its target value) and by the fact that output stability is only one (though important) consideration for government and voters (by contrast, monetary stability can be made the sole objective of the central bank). Hence, it is important to consider the case where the Treasury pursues its own agenda and sticks to its preferred value of the output target, y_T, rather than follow the optimal output target, y_T^*, that society assigns to it. Although in section 4 we restricted attention to the case $y_T = y_T^*$, Section 6.5 below considers both cases: i.e. $y_T = y_T^*$ and $y_T \neq y_T^*$.

We proceed as follows. First, we derive the optimal rational expectations precommitment solution in this more general setting (Proposition 6). In general, this solution is time-inconsistent and, therefore, requires a commitment technology. We then consider the *optimal delegation regime* (first considered in Section 4.1, above). If the Treasury follows the optimal output target (i.e. $y_T = y_T^*$), then the optimal delegation regime achieves the precommitment solution for all values of the parameters (Proposition 7). If, however, the Treasury cannot be given the optimal output target, and has its own agenda (i.e. $y_T \neq y_T^*$), then Section 6.7, below, shows that a 'near optimal' solution can still be achieved. What if the Treasury is not given an inflation target or does not care about inflation at all, but is willing to adopt the socially optimal output target? Section 6.8, below, shows that the optimal precommitment solution can still be achieved.

6.2 Description of the general model

The model is described by the following basic equations:

$$\text{Aggregate Demand}: y = \varphi f - \lambda\left(i - \pi^e\right) + \epsilon \tag{4.17}$$

$$\text{Aggregate Supply}: y = \mu\left(\pi - \pi^e\right) \tag{4.18}$$

$$\text{Society's Objective}: \ U_S = -\frac{1}{2}\alpha\pi^2 - \frac{1}{2}\beta\left(y - y_S\right)^2 - \frac{1}{2}\gamma f^2 \tag{4.19}$$

The parameters α, β, γ, φ, λ, μ are all strictly positive. φ and λ are a measure of the effectiveness of fiscal and monetary policy respectively in influencing aggregate demand and μ indicates the strength of inflation surprises in influencing aggregate supply. Finally, α, β, γ are the relative weights given to the various terms in the objective function. The state contingent values of the demand shock, ϵ, are:

$$\text{Bad State: } \epsilon_- = \rho x - (1-p)s \tag{4.20}$$

$$\text{Good State: } \epsilon_+ = \rho x + ps \tag{4.21}$$

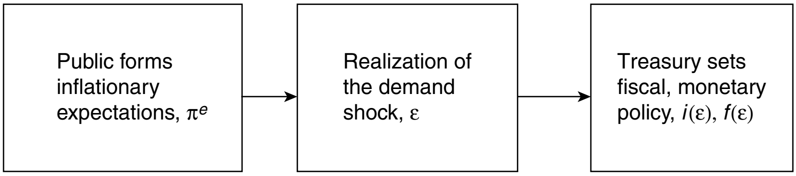

Public forms inflationary expectations, π^e
Realization of the demand shock, ε
Treasury sets fiscal, monetary policy, $i(\varepsilon)$, $f(\varepsilon)$

Figure 4.4 Sequence of moves when treasury controls *i, f*

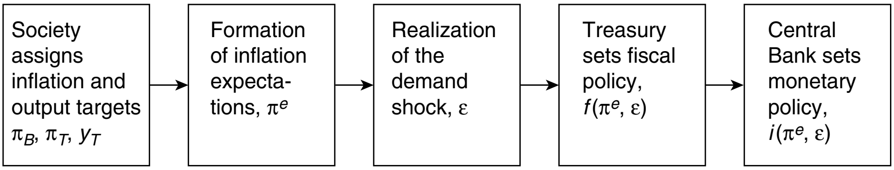

Society assigns inflation and output targets π_B, π_T, y_T
Formation of inflation expectations, π^e
Realization of the demand shock, ε
Treasury sets fiscal policy, $f(\pi^e, \varepsilon)$
Central Bank sets monetary policy, $i(\pi^e, \varepsilon)$

Figure 4.5 Sequence of moves in the optimal delegation regime

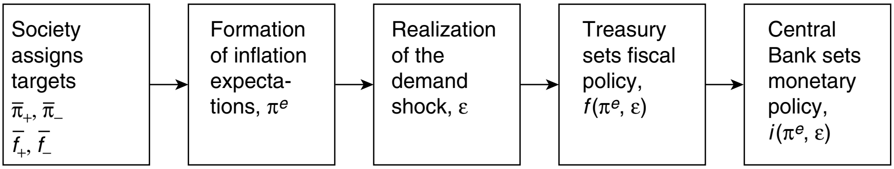

Society assigns targets $\bar{\pi}_+$, $\bar{\pi}_-$ $\bar{f}_+$, $\bar{f}_-$
Formation of inflation expectations, π^e
Realization of the demand shock, ε
Treasury sets fiscal policy, $f(\pi^e, \varepsilon)$
Central Bank sets monetary policy, $i(\pi^e, \varepsilon)$

Figure 4.6 Sequence of moves in the optimal delegation regime with a government budget constraint

where $0 < p < 1$, $s > 0$ and $0 \leq \rho < 1$. The variable x represents the previous period's shock and so ρ is a measure of the persistence in the shock. The second component in (4.20), (4.21) shows the innovation terms. With probability p the shock takes the value ϵ_- and with probability $1-p$ it takes a value ϵ_+. Hence $E[\epsilon|x] = p\epsilon_- + (1-p)\epsilon_+ = \rho x$ and so in the absence of the persistence term ($\rho = 0$), $E[\epsilon|x] = 0$ as in the model presented in Section 2.[29] Thus, if an economy is close to a liquidity trap, a negative shock can push the economy into it. Because of persistence, it may take the economy several periods to get out of the liquidity trap.

6.3 Sequence of moves and informational assumption

The sequence of moves under the regimes of precommitment, discretion and the optimal delegation are as in Figures 4.3, 4.4, 4.5 respectively, except that in any period, the realization of ϵ depends on the value of the of the shock in the previous period, x. We assume that formation of inflation expectations, π^e, and nominal wage contracts are signed after the observation of x but before the innovation (ps or $-(1-p)s$) is observed. However, the Treasury and Central

Bank have an informational advantage over the private sector in that they can set their instruments after the realization of the innovation part of the shock.[30] The main effect of this is to cause the optimal inflation and output targets to be state dependent (i.e., dependent on x). This is in line with the results of Eggertsson and Woodford (2003) who, however, consider only monetary policy.

6.4 Optimal solution

The optimal rational expectations precommitment solution, the analog of Proposition 1, is described below in Proposition 6. The intuition behind the results is similar to that behind Proposition 1 except that the magnitude of demand shocks in the past influence the state of the economy in the current period and so one needs to distinguish between three cases. Our main focus is on Case (b) where the economy is liquidity trapped in the bad state. The proof is derived analogously to that of Proposition 1 and, so, is omitted.

Proposition 6: ***(a)*** *If* $x < -ps\frac{\left(\alpha+\beta\mu^2\right)\left(\alpha\varphi^2+\gamma\lambda^2\right)}{\alpha\rho\left(\gamma\mu^2+\varphi^2\left(\alpha+\beta\mu^2\right)\right)}$ *then the economy is liquidity trapped in both states and the commitment solution is given by* $i_- = i_+ = 0$*,*

$$f_- = \varphi\left(\frac{\left(\alpha+\beta\mu^2\right)s\left(1-p\right)}{\gamma\mu^2+\varphi^2\left(\alpha+\beta\mu^2\right)} - \frac{\alpha\rho x}{\alpha\varphi^2+\gamma\lambda^2}\right) > 0$$

$$f_+ = -\varphi\left(\frac{\left(\alpha+\beta\mu^2\right)sp}{\gamma\mu^2+\varphi^2\left(\alpha+\beta\mu^2\right)} + \frac{\alpha\rho x}{\alpha\varphi^2+\gamma\lambda^2}\right) > 0$$

(b) *If* $-ps\frac{\left(\alpha+\beta\mu^2\right)\left(\alpha\varphi^2+\gamma\lambda^2\right)}{\alpha\rho\left(\gamma\mu^2+\varphi^2\left(\alpha+\beta\mu^2\right)\right)} \leq x < \left(1-p\right)\frac{s}{\rho}$ *then the economy is liquidity trapped in the bad state only and the commitment solution is given by* $i_- = f_+ = 0$*,*

$$f_- = \frac{\alpha\varphi\left(\alpha+\mu^2\beta\right)\left(\left(1-p\right)s-\rho x\right)}{\left(\alpha+\beta\mu^2\right)\left(\alpha\varphi^2+\gamma\lambda^2 p\right)+\alpha\gamma\mu^2\left(1-p\right)} > 0$$

$$i_+ = \frac{\left(\gamma\lambda^2+\alpha\varphi^2\right)\left(\alpha+\beta\mu^2\right)sp+\left(\beta\varphi^2\mu^2+\gamma\mu^2+\alpha\varphi^2\right)\alpha\rho x}{\lambda\left(\left(\alpha+\beta\mu^2\right)\left(\alpha\varphi^2+\gamma\lambda^2 p\right)+\alpha\gamma\mu^2\left(1-p\right)\right)} \geq 0$$

(c) *If* $x \geq \left(1-p\right)\frac{s}{\rho}$ *then the economy is liquidity trapped in neither state and the commitment solution is given by* $f_- = f_+ = 0$*,*

$$i_- = \frac{\rho x-\left(1-p\right)s}{\lambda} \geq 0$$

$$i_+ = \frac{\rho x+ps}{\lambda} > i_- \geq 0$$

Proposition 6 illustrates the evolution of the economy over time. Suppose that the economy is liquidity trapped in period t. How does it get out of a liquidity trap? Proposition 6 (b), (c) gives the conditions required on how big the shocks must be in period t so that in period $t+1$ the economy is not liquidity trapped in at least in one state of the world.[31]

6.5 The optimal delegation regime

In this section we examine the possibility of achieving the optimal precommitment solution through appropriate institutional design. Here we extend the *optimal delegation* framework of Section 4.1 (details are suppressed to avoid repetition) to incorporate the five extensions E1 through E5 stated at the beginning of Section 6. The Treasury's objective function is given by

$$U_T = -\frac{1}{2}\alpha\left(\pi - \pi_T\right)^2 - \frac{1}{2}\beta\left(y - y_T\right)^2 - \frac{1}{2}\gamma f^2 \qquad (4.22)$$

Note that the parameters α, β, γ are the same as in society's welfare function given in (4.19). Denote the optimal inflation target of the Central Bank by π_B^* and the optimal output and inflation targets of the Treasury by y_T^* and π_T^* respectively. Proposition 7, below, states the results under optimal delegation. As in Proposition 6, the magnitude of the demand shock in the previous period gives rise to three subcases, although we are primarily interested in Case (b). The proof is similar to that of Proposition 3, so it is omitted.

Proposition 7: ***(a)*** *Under the condition of Proposition 6(a), give the Central Bank any inflation target, π_B^*, that satisfies $\pi_B^* > \gamma\left(\frac{\mu s p}{\beta\varphi^2\mu^2+\alpha\varphi^2+\gamma\mu^2} + \frac{\lambda\rho(-x)}{\alpha\varphi^2+\gamma\lambda^2}\right)$ and give the Treasury any output and inflation target pair (y_T, π_T) that satisfy*

$$y_T\left(\pi_T\right) = k - \frac{\alpha}{\beta\mu}\pi_T \qquad (4.23)$$

where $k = \alpha\frac{(\lambda+\mu)\gamma\rho(-x)}{\beta\mu\left(\alpha\varphi^2+\lambda^2\gamma\right)}$. Then the solution under optimal delegation is the same as under precommitment, and given by Proposition 6(a).

(b) *Under the conditions of Proposition 6(b), give the Central Bank the inflation target*

$$\pi_B^* = \frac{\gamma\left(\beta\mu^2\lambda + \alpha\left(\lambda+\mu\right)\right)\left(s\left(1-p\right) - \rho x\right)p}{\left(\alpha+\mu^2\beta\right)\left(\alpha\varphi^2+\gamma\lambda^2 p\right) + \gamma\mu^2\alpha\left(1-p\right)} > 0 \qquad (4.24)$$

and give the Treasury any output and inflation target pair (y_T, π_T) that satisfies

$$y_T\left(\pi_T\right) = K - \frac{\alpha}{\beta\mu}\pi_T \qquad (4.25)$$

where $K = \frac{\alpha\gamma p}{\mu\beta}\frac{(\lambda+\mu)\left(\alpha+\mu^2\beta\right)(\epsilon(1-p)-\rho x)}{\left(\alpha+\beta\mu^2\right)\left(\alpha\varphi^2+\gamma\lambda^2 p\right)+\gamma\mu^2\alpha(1-p)}$. Then the solution under optimal delegation is the same as under precommitment and is given by Proposition 6(b). Furthermore, $\pi_+ = \pi_B^*$.

(c) *Under the condition of Proposition 6(c), give the Central Bank the inflation target $\pi_B^* = 0$. Then, for any output and inflation target pair (y_T, π_T) for the Treasury, the solution under optimal delegation is the same as under commitment and is given by Proposition 6(c). Furthermore, $\pi_+ = \pi_- = \pi_B^* = 0$.*

The intuition behind the optimality of this delegation regime is as in Section 4.1 above. If the economy is not liquidity trapped in any state of the world we are in the standard textbook case where delegation to an independent Central Bank achieves the precommitment solution. Proposition 7(c) deals with this case. Our main case of interest, however, is when the economy is liquidity trapped in the bad state only; this is stated in Proposition 7(b). Here, the inflation target of the Central Bank is uniquely determined while the Treasury's target pair y_T, π_T can be chosen from a *menu of contracts* that satisfy (4.25).

To explain the indeterminacy of y_T and π_T, note that the Treasury has *two* targets, y_T and π_T, but just *one* instrument, f. Hence, the best it can hope for is hit just one of these targets or, more generally, a linear combination of them. Maximizing society's expected welfare yields the optimal linear combination of y_T and π_T. This is given by (4.23) in the case of Proposition 7(a) and (4.25) in the case of Proposition 7(b). The negative slope signifies that a high output bias is needed to compensate a low inflation target for the Treasury.

What if the inflation targets of the Treasury and the Central Bank are identical? Corollary 2 describes the results when the economy is liquidity trapped in the bad state.

Corollary 2: *Under the conditions of Proposition 6(b), if $\pi_T = \pi_B^*$, then the optimal output target for the Treasury is*

$$y_T^* = \alpha\gamma p\mu^2 \frac{s(1-p) - \rho x}{\left(\alpha + \beta\mu^2\right)\left(\alpha\varphi^2 + \gamma\lambda^2 p\right) + \gamma\mu^2\alpha(1-p)} > 0. \qquad (4.26)$$

and the Treasury attains this target in the good state i.e. $y_+ = y_T^*$.

In Figure 4.7, the downward sloping line AA' is a graph of $y_T(\pi_T)$ defined in (4.23) or (4.25). The vertical line positioned at π_B^* reflects the inflation target

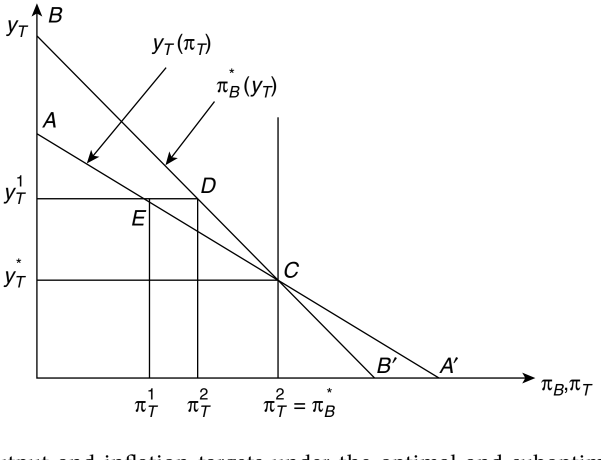


Figure 4.7 Output and inflation targets under the optimal and suboptimal delegation regimes

for the central bank given in 4.24. Ignore the downward sloping line BB' for the moment.

Proposition 7 then shows how the *optimal delegation regime* can achieve the *optimal precommitment solution* in the following two cases,

1. *The Treasury and the Central Bank can be given the same inflation target* Figure 4.7 shows that the optimal delegation solution is given by point C, where $\pi_B = \pi_T = \pi_B^*$ (given in (4.24)) and $y_T = y_T^*$ (given in (4.26)).
2. *The Treasury and the Central Bank are given distinct inflation targets* Figure 4.7 shows one possible solution. The Central Bank is given the uniquely determined inflation target i.e. $\pi_B = \pi_B^*$ (see (4.24)). The Treasury is given any output, inflation target along the line AA, for instance, corresponding to point E i.e. $(y_T, \pi_T) = \left(y_T^1, \pi_T^1\right)$.

6.6 Discretion

The results under discretion when we extend the basic model to extensions E1-E5 are similar to those stated in Proposition 2. The full set of results are given in Appendix-B; the method of proof is identical to that of Proposition 2, and is omitted. Denote by EU^{Disc}, the expected welfare level under discretion; we make use of it in Section 6.7 below.

6.7 Suboptimal delegation: Treasury follows its own agenda ($y_T \neq y_T^*$)

We now consider the case where the Treasury does not adopt the optimal output target (see discussion in Subsection 6.1.2 above); we call this regime '*suboptimal delegation*'. The output target y_T now represents the Treasury's own agenda and it refuses to accept the optimal output target, y_T^*. The objective function of the Treasury is given in (4.22). For pedagogical simplicity, we stick here to the more natural case where the inflation targets of the Treasury and the Central Bank are equal i.e. $\pi_B = \pi_T$.

Let $\pi_B^*(y_T)$ maximize society's expected welfare, given the output target, y_T, of the Treasury. For the general case in Section 6 the expression for $y_T\left(\pi_B^*\right)$ is too unwieldy, but for the simple model presented in Section 2 it is given by

$$y_T\left(\pi_B^*\right) = \frac{7}{4}a - \frac{11}{2}\pi_B^*$$

In Figure 4.7, the line BB' is a sketch of (the inverse of) $\pi_B^*(y_T)$. Any point on the line BB' gives the optimal inflation target for the Central Bank, $\pi_B^*(y_T)$, conditional on the Treasury's private, but not necessarily optimal, output target, y_T. which is steeper than the schedule $y_T(\pi_T)$ plotted as line AA'.

Suppose that the Treasury's output target is given by $y_T = y_T^1$. Then, at one possible suboptimal equilibrium $\pi_B = \pi_T = \pi_T^2$ while $y_T = y_T^1$ i.e. the Treasury's equilibrium targets are shown by the point D. Because point D is off the line AA, which plots the optimal menu of contracts for the Treasury, how well does the suboptimal delegation regime fare, relative to the optimal precommitment solution? Simulations, below, show that the performance of the *suboptimal delegation regime* is 'near optimal' and much better than *discretion*.

Denote the expected social welfare level under suboptimal delegation by EU_S^{SD}. The state contingent values of the policy variables in this case run into several pages, so we confine ourselves to reporting a representative sample of simulation results. Towards this end we define the following variables.

$q = EU_S^{Opt}/EU_S^{SD}$ is the expected welfare level under the optimal solution relative to the expected welfare under suboptimal delegation. Note that $0 < q \leq 1$ and $q = 1$ when $y_T = y_T^*$ (see Proposition 7).

$\omega = EU_S^{Disc}/EU_S^{SD}$ is the ratio of the expected welfare under discretion relative to that under suboptimal delegation. Note that $\omega > 0$ because the numerator and denominator are both negative.

$Q = \frac{EU_S^{SD} - EU_S^{Disc}}{EU_S^{Opt} - EU_S^{Disc}}$ is the ratio of the welfare loss under suboptimal delegation relative to that under the optimal solution when each is expressed as a difference from the expected welfare level under discretion. Hence, relative to the discretionary solution as a benchmark, this is the proportional loss to society in moving from the optimal solution to the suboptimal delegation solution. Note that $Q = 1$ for $y_T = y_T^*$ (see Proposition 7).

$o = y_S/y_T^*$ is the output target of society relative to the optimal output target given to the Treasury.

$t = y_T/y_T^*$ is the output target of the Treasury relative to the optimal output target given to it.

The feasible set of parameters belongs to a ten dimensional set. We give below simulations for a representative sample of parameters in Tables 4.1, 4.2 below. Tables 4.4 through 4.6 in Appendix-C give further simulation results to support our assertions. To simplify results, we focus on cases where the output targets of the Treasury and society coincide i.e. $y_T = y_S$ (and so $o = t$) and the inflation targets of the Treasury and the Central Bank also coincide i.e. $\pi_T = \pi_B$.

The main results of the simulations can be summarized as follows.

Proposition 8: *Even if the private agenda of the Treasury, i.e. y_T, is substantially different from the optimal output target, y_T^*, the expected welfare level under the suboptimal delegation solution is very close to the optimal precommitment solution i.e. q is very close to 1. Suppose that we start with the minimal institutional framework of the discretionary regime. Then moving to the institutional regime of suboptimal delegation recovers, for all parameter values that we have investigated, a very large percentage of the benefit that might accrue if one could move to the optimal solution i.e. Q is typically very close to 1.*

In Table 4.1, the economy is liquidity trapped in the bad state only. Even if the output target of the Treasury is up to 602.2 times higher than the optimal

Table 4.1 $p=\frac{1}{2},\ y_T=y_S=s,\ x=0$

α	φ	λ	β	γ	μ	q	ω	Q	$o=t$	$\pi_B^*(y_T)$
$\frac{10}{10}$	$\frac{10}{10}$	$\frac{1}{10}$	$\frac{1}{10}$	$\frac{10}{10}$	$\frac{1}{10}$	0.9999	1.007	0.9844	404.4	$0.045s$
$\frac{10}{10}$	$\frac{10}{10}$	$\frac{1}{10}$	$\frac{1}{10}$	$\frac{10}{10}$	$\frac{10}{10}$	0.9936	1.039	0.8589	6.422	$0.146s$
$\frac{10}{10}$	$\frac{10}{10}$	$\frac{10}{10}$	$\frac{1}{10}$	$\frac{10}{10}$	$\frac{1}{10}$	0.9999	1.270	0.9995	602.6	$0.178s$
$\frac{10}{10}$	$\frac{10}{10}$	$\frac{10}{10}$	$\frac{1}{10}$	$\frac{10}{10}$	$\frac{10}{10}$	0.9921	1.451	0.9828	8.6	$0.216s$
$\frac{1}{10}$	$\frac{10}{10}$	$\frac{1}{10}$	$\frac{1}{10}$	$\frac{10}{10}$	$\frac{1}{10}$	0.9989	1.067	0.9844	44.42	$0.406s$
$\frac{10}{10}$	$\frac{1}{10}$	$\frac{1}{10}$	$\frac{1}{10}$	$\frac{10}{10}$	$\frac{1}{10}$	1.0000	3.931	1.0000	8.006	$2.497s$
$\frac{1}{10}$	$\frac{1}{10}$	$\frac{1}{10}$	$\frac{1}{10}$	$\frac{10}{10}$	$\frac{10}{10}$	0.9999	1.006	0.9851	2.048	$0.585s$
$\frac{1}{10}$	$\frac{1}{10}$	$\frac{1}{10}$	$\frac{1}{10}$	$\frac{10}{10}$	$\frac{10}{10}$	0.9873	1.039	0.7521	2.84	$0.371s$

Table 4.2 $p=\frac{1}{50}, y_T=y_S=ps, x=-(1-p)s, \rho=\frac{9}{10}$

p	α	φ	λ	β	γ	μ	ω	π_B^*
$\frac{1}{50}$	$\frac{10}{10}$	$\frac{10}{10}$	$\frac{1}{10}$	$\frac{1}{10}$	$\frac{10}{10}$	$\frac{1}{10}$	1.0397	$0.17445s$
$\frac{1}{50}$	$\frac{10}{10}$	$\frac{10}{10}$	$\frac{1}{10}$	$\frac{1}{10}$	$\frac{10}{10}$	$\frac{10}{10}$	2.4680	$0.95859s$
$\frac{1}{50}$	$\frac{10}{10}$	$\frac{10}{10}$	$\frac{10}{10}$	$\frac{1}{10}$	$\frac{10}{10}$	$\frac{1}{10}$	2.4215	$0.4849s$
$\frac{1}{50}$	$\frac{1}{10}$	$\frac{10}{10}$	$\frac{1}{10}$	$\frac{1}{10}$	$\frac{10}{10}$	$\frac{1}{10}$	1.4805	$1.6016s$

output target (i.e. $o=t=602.2$), q and Q are still very close to 1. Tables 4–6, in the appendix, confirm these results for other parameter values. In Table 4.2, below, constructed under the conditions of Proposition 7(a), the economy is liquidity trapped in both states and there is a very high level of persistence in the demand shocks.

From Table 4.2, the social loss in the discretionary regime is, in some cases, twice that under suboptimal delegation.

6.8 What happens if the treasury does not have an inflation target?

Here we consider two alternative regimes. In both of these cases, the Central Bank is given an inflation target π_B, i.e., has the objective function given in (4.10) but the Treasury is not given an inflation target. We find that these regimes are able to achieve the precommitment solution.

6.8.1 The Treasury is an "output nutter"

If the Treasury is not given an inflation target, we call it an *output nutter.* Its objective function is then given by

$$U_T=-\frac{1}{2}\beta(y-y_T)^2-\frac{1}{2}\gamma f^2$$

Table 4.3 Treasury is an "output nutter" ($p=\frac{1}{50}, y_T = y_S = s \neq y_T^*, x=0$)

p	φ	λ	β	γ	μ	q	Q
$\frac{1}{50}$	$\frac{10}{10}$	$\frac{1}{10}$	$\frac{1}{10}$	$\frac{10}{10}$	$\frac{1}{10}$	.085849	−33829
$\frac{1}{50}$	$\frac{10}{10}$	$\frac{1}{10}$	$\frac{1}{10}$	$\frac{10}{10}$	$\frac{1}{10}$	0.95778	0.38282
$\frac{1}{50}$	$\frac{10}{10}$	$\frac{10}{10}$	$\frac{1}{10}$	$\frac{10}{10}$	$\frac{1}{10}$	0.20225	−3821.0
$\frac{1}{50}$	$\frac{10}{10}$	$\frac{10}{10}$	$\frac{1}{10}$	$\frac{10}{10}$	$\frac{10}{10}$	0.95911	0.2807

6.8.2 *The Treasury is a "reckless output nutter"*

If the Treasury cares neither about inflation nor the costs of fiscal policy we call it a *reckless output nutter*. Its objective function is then given by

$$U_T = -\frac{1}{2}(y - y_T)^2$$

We are interested in evaluating the performance of the alternative institutional regimes in which the Treasury does not care about inflation. Proposition 9, below, shows that the optimal precommitment solution can be achieved; the proof is identical to that of Proposition 3 and, so, is omitted.

Proposition 9: *Unless the economy is liquidity trapped in both states of the world, if the Treasury can be assigned an optimal output target y_T^* and the Central Bank is assigned an optimal inflation target, π_B^*, then the outcome in the "output nutter" and the "reckless output nutter" cases is identical to the precommitment regime.*

However, and unlike the suboptimal delegation regime, if the Treasury does not adopt the optimal output target, y_T^*, then the outcome can be very poor, and much worse than the outcome under discretion. Table 4.3 gives a sample of results for the "output nutter" case.

In this case, Q can take extreme negative values i.e. the output nutter regime turns out to be much worse than discretion; we summarize this result in the Proposition below.

Proposition 10: *If the Treasury is not assigned the optimal output target, y_T^*, then the performance of the "output nutter" and the "reckless output nutter" regimes can be very adverse and, possibly, much worse than the discretionary regime. In particular, if monetary policy is delegated to an independent central bank, with an optimal inflation target, while the Treasury retains discretion over fiscal policy, then the outcome can be poor and much worse than had the Treasury retained discretion over both monetary and fiscal policy.*

Proposition 10 indicates the serious consequences that can arise if the Treasury/government does not have the appropriate inflation or output targets even if it follows society's most preferred output target (note $y_T = y_S$ in Table 4.3). This has relevance for understanding the Japanese experience in which the fiscal authorities, as pointed out earlier, were not delegated with the optimally designed targets.

6.9 Summary

Proposition 7 and Corollary 2 establish that the *optimal delegation regime* (where the Bank has an optimal inflation target and the Treasury has optimal output and inflation targets) achieves the precommitment solution for all parameter values. Proposition 8 shows that performance of the *suboptimal delegation regime* (similar to the *optimal delegation regime*, except that the Treasury has its own output target) is near optimal, and much better than *discretion*, even when the Treasury deviates considerably from the optimal output target. Proposition 9 establishes that the *output nutter* and the *reckless output nutter* (in both cases the Bank and Treasury are given optimal inflation and output targets, respectively, but the Treasury is not given an inflation target) regimes also achieve the precommitment solution. However, Proposition 10 shows that the latter two regimes, unlike the *suboptimal delegation* regime, perform poorly, and can be much worse than *discretion*, if the Treasury deviates from the optimal output target. Thus, although giving the Treasury an inflation target as well as an output target is not necessary for optimality, it is necessary to achieve robustness. In particular, a hybrid system, where monetary policy is delegated to an independent central bank with an inflation target, but where the Treasury retains discretion over fiscal policy, can perform poorly and much worse than had the Treasury retained discretion over both monetary and fiscal policy.

7 Relation to the literature

The role of fiscal policy in theoretical models on the liquidity trap has not been adequately stressed despite this being Keynes's (1936) original solution to the problem. This is puzzling in light of the empirical evidence from Posen (1998), Kuttner and Posen (2001) which suggests that fiscal policy, when used in Japan, has been potent. The simulation exercises of Ball (2005) show that fiscal transfers equal to 6.6 percent of GDP could have ended Japan's output slump. There have been other suggestions in the literature, without a full theoretical model, that advocate fiscal policy in a liquidity trap. Bernanke (2002) recommends a broad based tax cut while Gertler (2003) recommends transitory fiscal policy. We consider fiscal policy explicitly in a Dixit and Lambertini (2003) framework when there is the possibility of a liquidity trap.

The theoretical literature has considered aspects of our *optimal delegation regime*, that achieves the precommitment solution. For instance, inflation targets have been suggested in Krugman (1998), Nishiyama (2003), and Iwamura et al. (2005). Other variants of monetary policy commitment have also been considered. Benhabib, Schmitt-Grohe and Uribe (2002) consider a commitment to switch from an interest rate rule to a money growth rate peg in a liquidity trap. Eggertsson and Woodford (2003) propose a commitment to adjust nominal interest rates to achieve a time varying price level target. Bernanke (2002) suggests a commitment to a buffer zone for the inflation rate. Svensson (2003) advocates a price level target (as part of a larger set of policies). However, none of these models allow for the possibility of strategic interaction between monetary and fiscal authorities nor jointly derive the optimal set of targets and instruments of the two policy making authorities.

Eggertsson (2006a, b) studies the liquidity trap within a new Keynesian stochastic general equilibrium model with a government budget constraint and explicit microfoundations. Eggertsson recommends abandonment of an independent central bank and a return to discretionary policy by a unitary monetary-fiscal authority. A debt financed fiscal expansion during a liquidity trap results, via the government budget constraint, in higher expectations of future inflation. Eggertsson shows that this solution is superior to either monetary policy alone or uncoordinated monetary and fiscal policy. However, as Eggertsson shows, even optimal discretion is inferior to the fully optimal rational expectations solution with commitment. Moreover, abandoning delegation of monetary policy to an independent central bank with a narrow mandate, in favor of a return to discretion, appears to be a retrograde step.

Dixit and Lambertini (2003) and Lambertini and Rovelli (2003) consider strategic interaction between fiscal and monetary authorities, but in the absence of a liquidity trap. Lambertini and Rovelli (2003) show that the equilibrium with the fiscal authority acting as leader is superior to the Nash equilibrium. Dixit and Lambertini (2003) show that this regime can achieve the optimal precommitment rational expectations solution.

One of the important lessons of our paper (see Figure 1.2 and Section 6) is that an optimally derived target for one policy maker while ignoring the incentives and constraints facing the other policy maker can lead to extremely poor outcomes; witness the last row in Figure 1.2.

Furthermore, the *optimal delegation regime* achieves the optimal mix between monetary and fiscal policy as we now explain. Theoretically, society could give a sufficiently high inflation target to the Central Bank which in turn generates sufficiently high inflation expectations so that the nominal interest rate never hits its zero floor. While this policy would always avoid the liquidity trap, it is not optimal because inflation is costly. Analogously it is not optimal to give the Treasury too high an output target because if a liquidity trap occurs, it would use the costly fiscal policy excessively. The optimal solution then is to have a mix of both i.e. some inflation outside a liquidity trap and some dependence on costly fiscal policy in a liquidity trap. The intuition is that if there were no liquidity trap, and the Treasury had its own agenda,[32] then it would undermine the Central Bank's monetary commitment. However, appropriate delegation of policy to the Treasury, far from undermining monetary commitment, gives it an incentive to engage in an 'appropriate' fiscal stimulus in a liquidity trap, where the independent Central Bank is ineffective.

8 Conclusions

In a liquidity trap, with nominal interest rates bound below by zero, an expectation of positive inflation is needed. This in turn needs a credible commitment to a future level of positive actual inflation. The credibility problem comes about because after the economy has escaped from the liquidity trap it is in the interest of all parties to renegotiate and reduce inflation. A forward looking private sector will anticipate this and expect low future inflation. With low expected future inflation, the real interest rate remains positive, keeping the economy in the liquidity trap; see for instance Krugman (1998).

The first best solution obtains when the rigidities that give rise to the liquidity trap are removed. But removal of these distortions is usually slow and difficult (witness the experience of Japan). In this case, macroeconomic policy can have an important role. Furthermore, the Japanese experience suggests that issues of strategic monetary fiscal policy interaction assume even greater importance in a liquidity trap.

In the solution considered here, society delegates monetary policy to an operationally independent Central Bank with an inflation target. Fiscal policy is delegated to the Treasury with inflation and output targets. Furthermore, the Treasury acts as a leader and the Central Bank is the follower. The required institutional arrangements are quite natural and are able to achieve the second best solution, namely, the best rational expectations precommitment solution. This institutional setting provides (1) the appropriate level of inflation and, hence, inflation expectations and (2) the optimal balance between monetary and fiscal policy. Even if the Treasury deviates considerably from the optimal output target, we find that the performance of this solution is still 'near optimal' and much better than the regime where the Treasury is given discretion over monetary and fiscal policy.

On the other hand, we find that the hybrid system where monetary policy is delegated to an independent central bank with an optimal inflation target, but where the Treasury retains discretion over fiscal policy, can perform badly and much worse than had the Treasury retained discretion over both fiscal and monetary policy. This is in line with the case when there is no liquidity trap considered by Dixit and Lambertini (2003, 1523, point 4): "Commitment achieves the second best only if it can be extended to both monetary and fiscal policy".

We have also considered a version of our model with an explicit government budget constraint. We found that the optimal solution can be achieved by giving appropriate inflation targets to the Central Bank and appropriate surplus/deficit targets to the Treasury,

9 Appendix-A: proofs of the main results

Generic Equilibrium: To save space, we carry out some calculations that are relevant to both Proposition 1 (Precommitment) and Proposition 2 (Discretion).

Substituting (4.4) and (4.5) into (4.6),

$$U_S = -\frac{1}{2}\left(f - i + \pi^e + \epsilon - y_S\right)^2 - \frac{1}{2}\left(f - i + 2\pi^e + \epsilon\right)^2 - f^2 \tag{4.27}$$

Since $f \gtrless 0$ and $i \geq 0$ the first order conditions are as follows.

$$\frac{\partial U}{\partial f} = y_S - 2\epsilon - 4f - 3\pi^e + 2i = 0;\ f \gtrless 0 \tag{4.28}$$

$$\frac{\partial U_S}{\partial i} = 2f - 2i + 3\pi^e + 2\epsilon - y_S \leq 0;\ i \geq 0 \quad \text{and} \quad i\frac{\partial U}{\partial i} = 0 \tag{4.29}$$

Since f is unrestricted, the optimal f satisfies $\frac{\partial U}{\partial f} = 0$, hence

$$f = \frac{1}{4}y_S - \frac{3}{4}\pi^e + \frac{1}{2}i - \frac{1}{2}\epsilon \tag{4.30}$$

Recall that values in the liquidity trap are distinguished by a '$-$' subscript and those in the complementary case by the '+' subscript. From (4.29), either $i \geq 0$ and $\frac{\partial U_S}{\partial i} = 0$ or $i = 0$ and $\frac{\partial U_F}{\partial i} < 0$, hence

$$i_+ = f_+ + \frac{3}{2}\pi^e - \frac{1}{2}y_S + a \quad \text{and} \quad f_+ + \frac{3}{2}\pi^e - \frac{1}{2}y_S + a \geq 0 \tag{4.31}$$

$$i_- = 0 \quad \text{and} \quad f_- + \frac{3}{2}\pi^e - \frac{1}{2}y_S - a < 0 \tag{4.32}$$

Substituting from (4.30), these two conditions can be restated as

$$i_+ = \frac{3}{2}\pi^e - \frac{1}{2}y_S + a \text{ and } 3\pi^e - y_S + 2a \geq 0 \tag{4.33}$$

$$i_- = 0 \text{ and } 3\pi^e - y_S - 2a < 0 \tag{4.34}$$

From (4.5), (4.30) (4.33), (4.34),

$$f_+ = 0 \tag{4.35}$$

$$f_- = \frac{1}{4}y_S - \frac{3}{4}\pi^e + \frac{1}{2}a \tag{4.36}$$

$$\pi_+ = \frac{1}{2}\left(y_S + \pi^e\right) \tag{4.37}$$

$$\pi_- = \frac{1}{4}y_S + \frac{5}{4}\pi^e - \frac{1}{2}a \quad \text{(liquidity trapped)} \tag{4.38}$$

This completes the description of the generic equilibrium. ■

Proof of Proposition 1 (Precommitment)
Since the two possible values of $\epsilon = -a$ and $\epsilon = a$ are equally probable, using (4.27) the expected social welfare is

$$E[U_S] = \frac{1}{2}\left(-\frac{1}{2}\left(f_+ - i_+ + \pi^e + a - y_S\right)^2 - \frac{1}{2}\left(f_+ - i_+ + 2\pi^e + a\right)^2 - f_+^2\right) + \frac{1}{2}\left(-\frac{1}{2}\left(f_- - i_- + \pi^e - a - y_S\right)^2 - \frac{1}{2}\left(f_- - i_- + 2\pi^e - a\right)^2 - f_-^2\right) \quad (4.39)$$

From (4.5), $\pi^e = f^e - i^e + 2\pi^e$, hence

$$\pi^e = \frac{1}{2}(i_+ + i_-) - \frac{1}{2}(f_+ + f_-) \quad (4.40)$$

Substituting (4.40) in (4.39) the expected social welfare is

$$E[U_S] = -\frac{1}{4}\left(\frac{1}{2}(f_+ - f_-) - \frac{1}{2}(i_+ - i_-) + a - y_S\right)^2 - \frac{1}{4}(i_- - f_- + a)^2 - \frac{1}{2}f_+^2 - \frac{1}{4}\left(\frac{1}{2}(i_+ - i_-) - \frac{1}{2}(f_+ - f_-) - a - y_S\right)^2 - \frac{1}{4}(i_+ - f_+ - a)^2 - \frac{1}{2}f_-^2 \quad (4.41)$$

We maximize $E[U_S]$ subject to $i_+ \geq 0$ and $i_- \geq 0$. Formally

$$\underset{\{f_-,f_+,i_-,i_+\}}{Max} \; E[U_S]$$

subject to

$$i_+ \geq 0,\; i_- \geq 0$$

Solving the first order conditions simultaneously, using the condition of rational expectations (4.40) and the equations for output and inflation in (4.4) and (4.5), one obtains the solution for the policy variables and the macroeconomic magnitudes reported in Proposition 1. ■

Proof of Proposition 2 (Discretion: Economy is liquidity trapped only under adverse demand conditions, $\epsilon = -a$)

Since $\epsilon = -a$ and $\epsilon = a$, each occur with probability $\frac{1}{2}$, the condition for rational expectations, using (4.37) and (4.38) gives:

$$\pi^e = \frac{1}{2}\left(-\frac{1}{2}a + \frac{1}{4}y_S + \frac{5}{4}\pi^e\right) + \frac{1}{2}\left(\frac{1}{2}\left(y_S + \pi^e\right)\right)$$

Hence the fixed point of π^e is readily found as

$$\pi^e = 3y_S - 2a \quad (4.42)$$

(4.33), (4.34) and (4.42) give

$$\frac{1}{2}a \leq y_S < a \quad (4.43)$$

which is the necessary and sufficient condition for this case to arise.

Substituting (4.42) in (4.33)–(4.36) gives the magnitudes of the policy instruments:

$$i_{-} = 0 \tag{4.44}$$

$$f_{-} = 2\left(a - y_S\right) > 0 \tag{4.45}$$

$$i_{+} = 4y_S - 2a \tag{4.46}$$

$$f_{+} = 0 \tag{4.47}$$

The magnitudes of output and inflation can now be found from (4.4), (4.5), (4.42), and (4.44)–(4.47):

$$y_{-} = y_S - a < 0 \tag{4.48}$$

$$\pi_{-} = 4y_S - 3a \tag{4.49}$$

$$y_{+} = a - y_S > 0 \tag{4.50}$$

$$\pi_{+} = 2y_S - a \tag{4.51}$$

The expected values (where expectations are taken over the demand shock ϵ) of i, f and y are given by

$$i^e = 2y_S - a$$

$$f^e = a - y_S > 0$$

$$y^e = 0$$

Hence, on average macroeconomic policy ensures that there are no deviations of output from the natural level ($y^e = 0$). To find the state-contingent levels of social welfare, substitute (4.45), (4.47), (4.48)–(4.51) into (4.6) then take expectations over the demand shock to get the expected social welfare

$$E\left[U_S^{Disc}\right] = 12ay_S - 8y_S^2 - 5a^2 \tag{4.52}$$

This completes the proof of the proposition. ■

Proof of Proposition 3 (Solution under the optimal delegation regime)

Monetary authority's reaction function

The monetary authority's reaction function can be found by maximizing U_B in (4.10). Since $i \geq 0$, the first order conditions for maximizing U_B are $\frac{\partial U_B}{\partial i} \leq 0$, $i \geq 0$, $i\frac{\partial U_B}{\partial i} = 0$. Using (4.5), this gives

$$i\left(f - i + 2\pi^e - \pi_B + \epsilon\right) \leq 0 \tag{4.53}$$

We start with the case where the economy is liquidity trapped in the the bad state *($\epsilon = -a$)* only. The other cases will be considered at the end.

The economy is in a liquidity trap ($\epsilon = -a$)

In this case, at $\epsilon = -a$ the interest rate can go no lower than zero. Using (4.53), $f_- + 2\pi^e - \pi_B - a < 0$, and so

$$i_- = 0 \tag{4.54}$$

The economy is not in a liquidity trap ($\epsilon = a$)

In this case, $i \geq 0$, hence (4.53) holds with equality. Solving out for i at $\epsilon = a$, gives

$$i_+ = f_+ + 2\pi^e - \pi_B + a \tag{4.55}$$

The state contingent reaction function of the monetary authority is given by (4.54) and (4.55).

Fiscal authority's reaction function

The Treasury now chooses its state contingent fiscal policy f to maximize the objective function (4.11) after observing π^e and ϵ and knowing that the state contingent reaction function of the monetary authority is given by (4.54) and (4.55).

Case-I: Liquidity trap ($\epsilon = -a$)

In this case, the subsequent monetary policy is $i_- = 0$, hence, using (4.4), (4.5), (4.11) and $\pi_T = \pi_B$, the government maximizes:

$$U_T^- = -\frac{1}{2}\left(f_- + \pi^e - a - y_T\right)^2 - \frac{1}{2}\left(f_- + 2\pi^e - a - \pi_B\right)^2 - f_-^2 \tag{4.56}$$

Maximizing U_T^- with respect to f_- gives

$$f_- = \frac{1}{2}a + \frac{1}{4}y_T + \frac{1}{4}\pi_B - \frac{3}{4}\pi^e \tag{4.57}$$

Case-II: No liquidity trap ($\epsilon = a$)

The subsequent monetary policy is given by (4.55), hence, using (4.4), (4.5), (4.11) and $\pi_T = \pi_B$, the government maximizes

$$U_T^+ = -\frac{1}{2}\left(\pi_B - \pi^e - y_T\right)^2 - f_+^2 \tag{4.58}$$

Maximizing U_T^+ with respect to f_+ gives

$$f_+ = 0 \tag{4.59}$$

The state contingent reaction function of the fiscal authority is given by (4.57) and (4.59) respectively.

Substituting the state contingent monetary and fiscal policy reaction functions in (4.4) and (4.5) one obtains

$$y_- = -\frac{1}{2}a + \frac{1}{4}y_T + \frac{1}{4}\pi_B + \frac{1}{4}\pi^e \tag{4.60}$$

$$\pi_- = -\frac{1}{2}a + \frac{1}{4}y_T + \frac{1}{4}\pi_B + \frac{5}{4}\pi^e \tag{4.61}$$

$$y_+ = \pi_B - \pi^e \tag{4.62}$$

$$\pi_+ = \pi_B \tag{4.63}$$

Calculation of expected inflation

Since the two states of the world are equally probable, π^e is simply a weighted average of inflation in (4.61) and (4.63) respectively

$$\pi^e = \frac{1}{3}y_T - \frac{2}{3}a + \frac{5}{3}\pi_B \tag{4.64}$$

Substituting π^e in (4.55), (4.57), (4.60)–(4.62), one obtains

$$f_- = a - \pi_B \tag{4.65}$$

$$y_- = -\frac{2}{3}a + \frac{1}{3}y_T + \frac{2}{3}\pi_B \tag{4.66}$$

$$\pi_- = -\frac{4}{3}a + \frac{2}{3}y_T + \frac{7}{3}\pi_B \tag{4.67}$$

$$i_+ = \frac{2}{3}y_T - \frac{1}{3}a + \frac{7}{3}\pi_B \tag{4.68}$$

$$y_+ = \frac{2}{3}a - \frac{1}{3}y_T - \frac{2}{3}\pi_B \tag{4.69}$$

Calculation of the optimal inflation target

Substituting (4.59), (4.63), (4.65), (4.66) (4.67), (4.69) in (4.11) the expected social welfare can be simplified and written as:

$$E\left[U_S^{SD}\right] = 3a\pi_B + \frac{2}{3}ay_T - \frac{7}{3}\pi_B^2 - \pi_B y_T - \frac{7}{6}a^2 - \frac{1}{6}y_T^2 - \frac{1}{2}y_S^2 \tag{4.70}$$

Maximizing $E\left[U_S^{SD}\right]$ in (4.70) with respect to π_B and y_T gives the following optimal inflation and output targets

$$\pi_B^* = \frac{3}{5}a \tag{4.71}$$

$$y_T^* = \frac{1}{5}a \tag{4.72}$$

Substituting (4.71) and (4.72) in (4.11) gives the final expression for expected social welfare in the Stackelberg delegation case

$$E\left[U_S^{SD}\right] = -\frac{1}{5}a^2 - \frac{1}{2}y_S^2 \tag{4.73}$$

Comparing with Proposition 1, we see that the inflation and output targets achieve the optimal solution, with the economy liquidity trapped in the bad state only. Hence, the two other cases, when the economy is never liquidity trapped and when the economy is liquidity trapped in both states, need not be considered; thus the proof is complete. ■

Proof of Proposition 4: Substitute from the behavioral equations (4.4) and (4.5) into society's social welfare function (4.6) to get U_S in terms of the policy instruments f and i and inflation expectation, π^e. Rewrite in terms of $f_\pm$ and $i_\pm$ according to subsection 2.3 and using (4.8) to get EU_S. Impose the government budget constraint (4.12). Maximize with respect to f, i_+, i_- taking account of the non-negativity constraints $i_+ \geq 0$, $i_- \geq 0$. This yields Proposition 4. ■

Proof of Proposition 5: Set the Treasury the fiscal targets $f_T^- = \frac{1}{4}a$ and $f_T^+ = -\frac{1}{4}a$. Then maximizing the Treasury's objective functions, (4.14) and (4.13), clearly gives $f_- = \frac{1}{4}a$ and $f_+ = -\frac{1}{4}a$. Give the Central Bank the inflation targets $\pi_B^- = \frac{1}{4}a$ and $B_B^+ = \frac{3}{4}a$. Then, clearly, $\pi^e = \frac{1}{2}a$. Using these, and (4.5), gives $\pi_- = \frac{1}{4}a - i_-$ and $\pi_+ = \frac{7}{4}a - i_+$. Substituting from these into the Central Banks's objective functions, (4.16) and (4.15), gives $U_B^- = -\frac{1}{2}(i_-)^2$ and $U_B^+ = -\frac{1}{2}(i_+ - a)^2$. These are clearly maximized at $i_- = 0$ and $i_+ = a$. Finally, use (4.4) and (4.5) to complete the proof. ■

9.1 Appendix B: the discretionary regime in the general case

Proposition 11: ***(a)*** *Let* $\sigma = \text{signum}\left(\alpha\varphi^2 - \gamma\lambda\mu\right)$. *If*

$$\sigma x < -\frac{\sigma}{\alpha\rho}\left(\beta\lambda\mu y_S + \frac{\left(\alpha\varphi^2 - \mu\lambda\gamma\right)\left(\alpha + \beta\mu^2\right)sp}{\gamma\mu^2 + \varphi^2\left(\alpha + \beta\mu^2\right)}\right)$$

then the economy is liquidity trapped in both states and the solution under discretion is given by $i_- = i_+ = 0$

$$f_- = \varphi\left(\frac{\left(\alpha + \beta\mu^2\right)s(1-p)}{\gamma\mu^2 + \varphi^2\left(\alpha + \beta\mu^2\right)} - \frac{\alpha\rho x + \beta\lambda\mu y_S}{\alpha\varphi^2 - \gamma\lambda\mu}\right) > 0$$

$$f_+ = \varphi\left(-\frac{\left(\alpha + \beta\mu^2\right)sp}{\gamma\mu^2 + \varphi^2\left(\alpha + \beta\mu^2\right)} - \frac{\alpha\rho x + \beta\lambda\mu y_S}{\alpha\varphi^2 - \mu\lambda\gamma}\right) > 0$$

(b) *Let* $\sigma = \text{signum}\left(\left(\gamma\lambda\mu p - \alpha\varphi^2\right)\left(\alpha + \beta\mu^2\right) - \alpha\gamma\mu^2(1-p)\right)$. *If*

$$-\frac{\sigma}{\alpha\rho}\left(\alpha sp + \beta\lambda\mu y_S - \alpha s\right) < \sigma x \leq -\frac{\sigma}{\alpha\rho}\left(\beta\lambda\mu y_S + \frac{\left(\alpha\varphi^2 - \gamma\lambda\mu\right)\left(\alpha + \beta\mu^2\right)sp}{\gamma\mu^2 + \varphi^2\left(\alpha + \beta\mu^2\right)}\right)$$

Table 4.4 $p=\frac{1}{50}, y_T=y_S=s, x=0$

α	φ	λ	β	γ	μ	q	ω	Q	$o=t$	$\pi_B^* \times 10^2$
$\frac{10}{10}$	$\frac{10}{10}$	$\frac{1}{10}$	$\frac{1}{10}$	$\frac{10}{10}$	$\frac{1}{10}$	1.0000	1.000	0.9483	5158.2	$0.037s$
$\frac{10}{10}$	$\frac{10}{10}$	$\frac{1}{10}$	$\frac{1}{10}$	$\frac{10}{10}$	$\frac{10}{10}$	0.9992	1.070	0.9882	106.13	$0.095s$
$\frac{10}{10}$	$\frac{10}{10}$	$\frac{10}{10}$	$\frac{1}{10}$	$\frac{10}{10}$	$\frac{1}{10}$	1.0000	1.001	0.9842	5259.3	$2.073s$
$\frac{10}{10}$	$\frac{10}{10}$	$\frac{10}{10}$	$\frac{1}{10}$	$\frac{10}{10}$	$\frac{10}{10}$	0.9992	1.058	0.9858	107.24	$1.863s$
$\frac{1}{10}$	$\frac{10}{10}$	$\frac{1}{10}$	$\frac{1}{10}$	$\frac{10}{10}$	$\frac{1}{10}$	0.9998	1.003	0.9576	566.34	$3.369s$
$\frac{10}{10}$	$\frac{1}{10}$	$\frac{1}{10}$	$\frac{1}{10}$	$\frac{10}{10}$	$\frac{1}{10}$	1.0000	1.034	1.0000	102.09	$1.959s$
$\frac{1}{10}$	$\frac{1}{10}$	$\frac{1}{10}$	$\frac{1}{10}$	$\frac{10}{10}$	$\frac{10}{10}$	1.0000	1.887	1.0000	50.122	$2.392s$
$\frac{1}{10}$	$\frac{1}{10}$	$\frac{1}{10}$	$\frac{1}{10}$	$\frac{10}{10}$	$\frac{10}{10}$	0.9984	1.902	0.9983	60.224	$1.825s$

then the economy is liquidity trapped in the bad state only and the solution under discretion is given by $i_- = f_+ = 0$

$$f_- = \varphi\frac{\left(\alpha+\beta\mu^2\right)\beta\lambda\mu y_S + \left(\alpha+\beta\mu^2\right)\alpha p x - \alpha s\left(\alpha+\mu^2\beta\right)(1-p)}{\left(\alpha\mu+\lambda\left(\alpha+\beta\mu^2\right)\right)\gamma\mu p - \alpha\left(\gamma\mu^2+\varphi^2\left(\alpha+\beta\mu^2\right)\right)} > 0$$

$$i_+ = \frac{-\left(\alpha p x+\beta\lambda\mu y_S\right)\left(\gamma\mu^2+\varphi^2\left(\alpha+\beta\mu^2\right)\right)+\left(\alpha+\mu^2\beta\right)\left(\gamma\lambda\mu-\alpha\varphi^2\right)sp}{\lambda\left(\left(\gamma\lambda\mu p-\alpha\varphi^2\right)\left(\alpha+\beta\mu^2\right)-\gamma\mu^2\alpha(1-p)\right)} \geq 0.$$

(c) If $x \geq \frac{\alpha s(1-p)-\lambda\beta\mu y_S}{\alpha p}$ *then the economy is liquidity trapped in neither state and the solution under discretion is given by* $f_- = f_+ = 0$*,*

$$i_- = \frac{\alpha p x+\alpha s p+\beta\lambda\mu y_S-\alpha s}{\alpha\lambda} \geq 0$$

$$i_+ = \frac{\alpha p x+\alpha s p+\beta\lambda\mu y_S}{\alpha\lambda} > i_- \geq 0$$

9.2 Appendix C: further simulation results

Tables 4.4, 4.5, 4.6 report the most interesting case: the economy is liquidity trapped in the bad state only.

Table 4.4 below confirms results similar to those in Table 4.1 when the probability of falling into the liquidity trap is very remote i.e. $p=\frac{1}{50}$.

In Table 4.4, even if the output target of the Treasury, y_T, is 5158.2 times that of the optimal output target, y_T^*, results R1 and R2 above still hold. Tables 4.5 and 4.6 below confirm the two main results, R1 and R2, for much smaller output targets of the Treasury $y_T=y_S=ps$ when the probability of falling into the liquidity trap takes a high and a low value respectively.

Table 4.5 $p=\frac{1}{2}, y_T=y_S=ps, x=0$

α	φ	λ	β	γ	μ	q	ω	Q	$o=t$	π_B^*
$\frac{10}{10}$	$\frac{10}{10}$	$\frac{1}{10}$	$\frac{1}{10}$	$\frac{10}{10}$	$\frac{1}{10}$	1.0000	1.01	0.9969	202.2	$0.01s$
$\frac{10}{10}$	$\frac{10}{10}$	$\frac{1}{10}$	$\frac{1}{10}$	$\frac{10}{10}$	$\frac{10}{10}$	0.9982	1.18	0.9904	3.211	$0.16s$
$\frac{10}{10}$	$\frac{10}{10}$	$\frac{10}{10}$	$\frac{1}{10}$	$\frac{10}{10}$	$\frac{1}{10}$	0.9999	1.48	0.9999	301.3	$0.18s$
$\frac{10}{10}$	$\frac{10}{10}$	$\frac{10}{10}$	$\frac{1}{10}$	$\frac{10}{10}$	$\frac{10}{10}$	0.9973	2.48	0.9982	4.3	$0.23s$
$\frac{1}{10}$	$\frac{10}{10}$	$\frac{1}{10}$	$\frac{1}{10}$	$\frac{10}{10}$	$\frac{1}{10}$	0.9996	1.14	0.9971	22.21	$0.43s$
$\frac{10}{10}$	$\frac{1}{10}$	$\frac{1}{10}$	$\frac{1}{10}$	$\frac{10}{10}$	$\frac{1}{10}$	1.0000	3.98	1.0000	4.003	$2.50s$
$\frac{1}{10}$	$\frac{10}{10}$	$\frac{1}{10}$	$\frac{1}{10}$	$\frac{10}{10}$	$\frac{10}{10}$	0.9985	1.17	0.9913	1.42	$0.41s$

Table 4.6 $p=\frac{1}{50}, y_T=y_S=ps, x=0$

α	φ	λ	β	γ	μ	q	ω	Q	$o=t$	$\pi_B^* \times 10^2$
$\frac{10}{10}$	$\frac{10}{10}$	$\frac{1}{10}$	$\frac{1}{10}$	$\frac{10}{10}$	$\frac{1}{10}$	1.0	1.001	1.0010	11.33	$0.3875s$
$\frac{10}{10}$	$\frac{10}{10}$	$\frac{1}{10}$	$\frac{1}{10}$	$\frac{10}{10}$	$\frac{10}{10}$	1.0	1.009	0.9999	2.123	$1.045s$
$\frac{10}{10}$	$\frac{10}{10}$	$\frac{10}{10}$	$\frac{1}{10}$	$\frac{10}{10}$	$\frac{1}{10}$	1.0	1.024	1.0000	105.19	$2.093s$
$\frac{10}{10}$	$\frac{10}{10}$	$\frac{10}{10}$	$\frac{1}{10}$	$\frac{10}{10}$	$\frac{10}{10}$	1.0	1.035	1.0000	2.145	$1.957s$
$\frac{1}{10}$	$\frac{10}{10}$	$\frac{1}{10}$	$\frac{1}{10}$	$\frac{10}{10}$	$\frac{1}{10}$	1.0	1.006	0.9999	11.327	$3.546s$
$\frac{10}{10}$	$\frac{1}{10}$	$\frac{1}{10}$	$\frac{1}{10}$	$\frac{10}{10}$	$\frac{1}{10}$	1.0	1.041	1.0000	2.042	$1.960s$
$\frac{1}{10}$	$\frac{10}{10}$	$\frac{1}{10}$	$\frac{1}{10}$	$\frac{10}{10}$	$\frac{10}{10}$	1.0	1.008	0.9999	1.205	$1.992s$

Notes

1. Average inflation rates in successive decades from the 1950's on to the current decade show a declining trend; see table 1 in Svensson (2003).
2. High unemployment is an obvious fallout of a liquidity trap. An increase in the real value of private debt has further adverse consequences particularly for the financial sector. An increase in the real public debt creates a difficult problem for the government to increase taxes to balance its books on the one hand but risk getting mired deeper into a recession on the other.
3. The real interest rate is given by $r = i - \pi^e$ where i is the nominal interest rate and π^e is expected inflation. In a liquidity trap, $i = 0$ and typically $\pi^e < 0$, hence $r > 0$. To expand economic activity, the government needs to lower r; one possible solution is to generate positive inflationary expectations.
4. Variants of the devaluation approach can be found in McCallum (2000) and Svensson (2003). There are several potential problems with the devaluation option. First, calibrated models show that the magnitude of the devaluation required to get out of the liquidity trap might be too high. Second, using the uncovered interest rate parity condition when the domestic interest rate is zero, the expected appreciation of the home currency is fully locked-in by the foreign interest rate. Third, current devaluation will generate expectations of future appreciation of currency when the economy moves out of the liquidity trap, generating counter flows that frustrate attempts to devalue.

Fourth, devaluations may bring about competitive devaluations or retaliations in the form of other barriers to trade.

5. In a liquidity trap, zero nominal interest rates make bonds and money perfect substitutes. Hence, it might be difficult to engineer a price level increase. Furthermore, increases in money supply, suggested, for instance, in Clouse et al. (2003) and in Orphanides and Wieland (2000), for a long enough period that exceeds the duration of the liquidity trap, creates problems of credibility. While short term interest rates might be zero, long term interest rates might be strictly positive (this has been true of Japan during its deflationary experience). Hence, several authors such as Bernanke (2002) and Auerbach and Obstfeld (2005) have suggested open market operations in long term bonds. However, moving the long run yield curve on securities is confounded by the presence of the risk premium term whose behavior in a liquidity trap is not well known. A carry tax on money, suggested by Buiter and Panigirtzoglu (2003), works in theory but substantial practical problems of implementation are likely.
6. Central bank independence has other benefits. For example, it shields monetary policy from political interference and allows the delegation of policy to the most competent experts etc.
7. There are clearly other relevant issues in the Japanese experience such as the ineffectiveness of monetary policy that we do not touch on here; see Blinder (2000).
8. Examples are Krugman (1998), Eggerston and Woodford (2003), Nishiyama (2003), Clouse et al. (2003), Buiter- Panigirtzoglou (2003), and Auerbach and Obstfeld (2005). Ball (2005) considers fiscal policy alone.
9. Examples include (1) monetary and fiscal policy in Benhabib, Schmitt-Grohe and Uribe (2002), Iwamura et al. (2005) and (2) monetary and exchange rate policy in Orphanides and Wieland (2000), McCallum (2000) and Svensson (2003). Bernanke (2002) considers both monetary and fiscal policy but there is no theoretical analysis.
10. In the first group are Krugman (1998), Eggertsson and Woodford (2003), Orphanides and Wieland (2000), McCallum (2000), and Svensson (2003). In the second group are Benhabib, Schmitt-Grohe and Uribe (2002), Nishiyama (2003), Clouse et al. (2003), Buiter-Panigirtzoglou (2003), and Auerbach and Obstfeld (2005).
11. A dental analogy might be appropriate here. Tooth decay can be prevented by extracting all the child's teeth. But, normally, the optimal policy is not to extract; tooth decay then occurs with some probability.
12. In the first group are Krugman (1998), Eggertsson and Woodford (2003), Benhabib, Schmitt-Grohe and Uribe (2002), Shin-Ichi (2003), Clouse et al. (2003), and Buiter and Panigirtzoglou (2003). In the second group are papers by Ball (2005), and Auerbach and Obstfeld (2005). Finally there are papers that touch on both ex-ante and ex-post issues, for instance, Orphanides and Wieland (2000), McCallum (2000), Bernanke (2002), and Svensson (2003).
13. By optimality or near optimality we mean regimes that help us to attain or get very close to the optimal rational expectations (or pre-commitment) solution.
14. See, for example, Romer (2006, chapter 10) and Walsh (2003, chapter 8).
15. To be more precise, f is the *stabilization* component of fiscal policy (which varies over the business cycle). *Total* fiscal policy is then $F = f_0 + f$, where f_0 is *fixed* and chosen so that $F^e = f_0 + f^e = 0$, so that the government budget constraint holds on average. This is discussed further in section 5, below.
16. In principal these alternative modes of finance need not be equivalent. However, in the context of a liquidity trap, Ball (2005) shows that there are no long run differences arising from these alternative modes of finance.

17. The following formulation might appear even more plausible

$$AD : y_t = f_t - \left(i_t - \pi^e_{t+1}\right) + \epsilon_t$$

$$AS : y_t = \pi_t - \pi^e_t$$

where $\pi^e_t = E_{t-1}\pi_t$ and $\pi^e_{t+1} = E_t\pi_{t+1}$. However, in our model, the private sector has to make its decision before the realization of the demand shock ϵ_t. Hence, in the aggregate demand curve, it has to forecast π^e_{t+1} at time $t-1$. But $E_{t-1}\pi^e_{t+1} = E_{t-1}\left(E_t\pi_{t+1}\right) = E_{t-1}\left(\pi_{t+1}\right) = E_{t-1}\left(\pi_t\right) = \pi^e_t$. While this is true in our model, it is not true more generally.

18. The modern literature on the liquidity trap stresses demand shocks as major contributory factors. We could also consider supply shocks. The main difference is as follows. A sufficiently *negative* demand shock will push the economy into a liquidity trap. On the other hand, a sufficiently *positive* supply shock will also create a liquidity trap. In either case, the real interest rate fails to drop sufficiently to match demand with supply. Hence our framework can be easily extended to incorporate supply, as well as demand, shocks.

19. However, our model has the following differences from Dixit-Lambertini. (1) We normalize the natural rate of output to zero, hence, the additive shock ϵ (in (4.1) or in (4.4)) can also be interpreted as a shock to the natural rate of output. (2) Our model has the New Keynesian feature that expected inflation, π^e, also affects actual inflation, π. (3) Our stochastic structure allows persistence (see section 6 below). While there is no persistence in Dixit-Lambertini, they allow all parameters to be stochastic, hence, considering the possibility of non-additive shocks. (4) In our model a fiscal impulse acts on the demand side, creating greater output and inflation. However, in Dixit-Lambertini fiscal policy works on the supply side and takes the form of a subsidy to imperfectly competitive firms that increases output but reduces prices.

20. Most dynamic structural models used in the analysis of a liquidity trap are forward looking New Keynesian models. Gertler (2003), Mankiw (2001) note dissatisfaction with this model in terms of its inability to explain persistence in the data. Recent work, for instance, Ruud and Whelan (2006), casts doubt even on the hybrid variant proposed by Gali and Gertler (1999). Of course, similar criticisms apply to the version of our model microfounded along the lines of Dixit and Lambertini (2003). Thus, all current macroeconomic models lack satisfactory microfoundations.

21. The microfoundations for this in Dixit and Lambertini (2000, 2003) rest on the presence of monopolistic competition. Monopoly power in the product market reduces output below the efficient level, hence, giving policy makers an incentive to raise output. There are also a large number of other well known reasons for (4.7) but the ultimate cause, argue Alesina and Tabellini (1987), is the absence of non-distortionary taxes. For if they were available then other market failures could be corrected.

22. Fiscal policy is typically more cumbersome to alter, on account of the cost of changing it (balanced budget requirements, lobby groups etc.). Indeed the 'monetary policy committee' in the UK or the Fed in the USA meet on a regular basis to make decisions on the interest rate while changes to the tax rates are much less frequent.

23. Strictly speaking, for our qualitative results to hold, we only require that fiscal policy be relatively more expensive than the (possibly strictly positive) cost of using monetary policy. Normalizing the cost of using monetary policy to zero, however, ensures greater tractability and transparency of the results.

24. Strictly speaking, this is a second best solution. The first best obtains if the imperfections responsible for the liquidity trap are removed. It is variously referred to as the 'precommitment solution', the 'optimal rational expectations solution', the 'second best solution' or simply the 'optimal solution'.

25. Recall that f refers only to the stabilization component of fiscal policy, hence, $f_+ = 0$ is consistent with a strictly positive level of government expenditure on other items such as redistribution etc.
26. We conjecture that the combination of rigid wages-prices and a flexible nominal interest rate has the effect that the real interest rate, $i - \pi^e$, overshoots so as to equilibrate the economy.
27. The full set of results under discretion is given in Appendix-B.
28. As stressed by Eggertsson and Woodford (2003), failure to meet the inflation target in the liquidity trap does not signify failure of policy. A similar remark can be made with respect to the output target.
29. In more standard but less convenient notation, $x_t = \rho x_{t-1} + z_t$, where $z_t = -(1-p)s$ with probability p and $z_t = ps$ with probability $1-p$.
30. This is the standard assumption in the time-inconsistency literature.
31. This might not be a bad descriptor of the actual occurrence of a liquidity trap given the deep reservations expressed about the efficacy of most macroeconomic policies; see Blinder (2000) for an excellent survey.
32. In Dixit and Lambertini (2003) the Treasury never has its own agenda and fully internalizes society's social welfare function.

References

[1] Alesina, A. and Tabellini, G. (1987). Rules and discretion with non-coordinated monetary and fiscal policies. *Economic Inquiry,* 619–630.

[2] Auerbach, A. J. and Obstfeld, Maurice. (2005). The case for open market operations in a liquidity trap. *American Economic Review*, 95, 110–137.

[3] Ball, L. (2005). Fiscal remedies for Japan's slump. *NBER Working Paper 11374*.

[4] Benhabib, J. Schmitt-Grohe, S., and Uribe, M. (2002). Avoiding liquidity traps. *Journal of Political Economy*, 110, 535–563.

[5] Bernanke, Ben S. (2002). Deflation: Making sure that it does not happen here. *Speech on November 21, Federal Reserve Board.*

[6] Blanchard, O., Dell'Ariccia, G., and Mauro, P. (2010). Rethinking macroeconomic policy. *IMF Staff Position Note, SPN/10/03*.

[7] Blinder, A. (2000). Monetary policy at the zero lower bound: Balancing the risks. *Journal of Money Credit and Banking*, 32, 1093–1099.

[8] Buiter, W. H. and Panigirtzoglou, N. (2003). Overcoming the zero bound on nominal interest rates with negative interest on currency: Gesell's solution. *The Economic Journal*, 113, 723–746.

[9] Clouse, J., Henderson, D., Orphanides, A., Small, D., and Tinsley, P. (2003). Monetary policy when the nominal short term interest rate is zero. *Topics in Macroeconomics, www.bepress.com*.

[10] Dhami, S. and al-Nowaihi (2011). Optimal institutional design when there is a zero lower bound on interest rates. *Oxford Economic Papers*, 63 (4), 700–721.

[11] Dixit, A. and Lambertini, L. (2000). Fiscal discretion destroys monetary commitment. *mimeo. Princeton University.*

[12] Dixit, A. and Lambertini, L. (2003). Interactions of commitment and discretion in monetary and fiscal policies. *American Economic Review*, 93, 1522–1542.

[13] Eggertsson, G. (2006a). The deflationary bias and committing to being irresponsible. *Journal of Money, Credit and Banking*, 38(2).

[14] Eggertsson, G. (2006b). Fiscal multipliers and policy coordination *Federal Reserve Bank of New York, Staff Report no. 241*.

[15] Eggertsson, G. and Woodford, M. (2003). The zero bound on interest rates and optimal monetary policy. *Brookings Papers On Economic Activity*, 1, 139–211.

[16] Gali, J. and Gertler, M. (1999). Inflation dynamics: A structural econometric analysis. *Journal of Monetary Economics*, 44, 195–222.

[17] Gertler, M. (2003). Comments on: The zero lower bound on interest rates and optimal monetary policy. Paper prepared for the *Brookings Panel on Economic Activity*, 1, 219–227.

[18] Iwamura, M., Kudo, T. and Watanabe, T. (2005). Monetary and fiscal policy in a liquidity trap: The Japanese experience 1999–2004. *NBER Working Paper No. W11151*. Available at SSRN: http://ssrn.com/abstract=669450.

[19] Keynes, M. (1936). *The general theory of employment, interest and money*. Macmillan: London.

[20] Krugman, P. (1998) Its baaack! Japan's slump and the return of the liquidity trap. *Brookings Papers on Economic Activity*, 2, 137–187.

[21] Krugman, P. (1999). Thinking about the liquidity trap. *mimeo., http://web.mit.edu/Krugman/www/trioshrt.html*.

[22] Kuttner, K. N. and Posen, A. (2001). The Great Recession: Lessons for macroeconomic policy from Japan. *Brookings Papers on Economic Activity*, 2001, 93–185.

[23] Lambertini, L. and Rovelli, R. (2003). Monetary and fiscal policy coordination and macroeconomic stabilization: A theoretical analysis. *Available at SSRN: http://ssrn.com/abstract=380322* or *DOI: 10.2139/ssrn.380322*.

[24] Mankiw, N. G. (2001). The inexorable and mysterious trade-off between inflation and unemployment. *Economic Journal*, 111, 45–61.

[25] McCallum, Bennet T. (2000). Theoretical analysis regarding a zero lower bound on nominal interest rates. *Journal of Money Credit and Banking*, 32, 870–904.

[26] Nishiyama, Shin-Ichi. (2003). Inflation target as a buffer against liquidity trap. *Institute for Monetary and Economic Studies*, Discussion Paper No. 2003-E-8.

[27] Orphanides, A. and Wieland, V. (2000). Efficient monetary policy design near price stability. *Journal of the Japanese and International Economies*, 14, 327–365.

[28] Rogoff, K. (1985). The optimal commitment to an intermediate monetary target. *Quarterly Journal of Economics*, 100 (4), 1169–1189.

[29] Romer, D. (2006). *Advanced Macroeconomics*, 3rd edition, McGraw-Hill/Irwin.

[30] Rotemberg and Woodford. (1999). Interest rate rules in an estimated sticky price model in Monetary Policy Rules. In John B. Taylor, ed., *Monetary Policy Rules*, University of Chicago Press, 57–126.

[31] Ruud, J. and Whelan, K. (2006). Can rational expectations sticky-price models explain inflation dynamics? *American Economic Review*, 96, 303–320.

[32] Posen, A. S. (1998). *Restoring Japan's Economic Growth*. Institute for International Economics, Washington DC.

[33] Sims, C. A. (2003). Implications of rational inattention. *Journal of Monetary Economics*, 50, 665–690.

[34] Svensson, Lars E. O. (2003). Escaping from a liquidity trap and deflation: The foolproof way and others. *Journal of Economic Perspectives*, 17, 145–166.

[35] Walsh, C. E. (2003). *Monetary Theory and Policy*, 2nd edition. The MIT Press.

[36] Woodford, M. (2005). Central bank communication and policy effectiveness. Presented at the Federal Reserve Bank of Kansas City Symposium, "The Greenspan Era: Lessons for the Future" Jackson Hole, Wyoming, 25–27 August 2005.

5
Analyzing Bank Efficiency: Are "Too-Big-to-Fail" Banks Efficient?

Hulusi Inanoglu, Michael Jacobs, Jr., Junrong Liu and Robin Sickles

1 Introduction

The recent financial crisis has given rise to a re-examination by regulators and academics of the conventional wisdom regarding the implications of the spectacular growth of the financial sector of the economy. In the pre-crisis era, there was a widespread common wisdom that "bigger is better." The arguments underpinning this view ranged from potential economies of scale and scope, to a better competitive stance at the international level. However, in the post-crisis world the common wisdom has been altered somewhat as large banks have come to be viewed as problematic for policy makers and regulators, for various reasons. One reason often given is that economic agents who are insured have the incentive to take on too much ex ante risk; also known as the moral hazard problem. Second, there is the "too-big-to-fail" problem: the fear that large and interconnected financial institutions may become a source of systemic risk if allowed to go out of business, especially in a "disorderly" fashion (Bernanke (2009)). Support for or against large banking institutions turns on the central issue of whether or not efficiencies of scale and scope are economically and statistically significant and are positively associated with bank size. If they are positively associated with bank size then the expected benefits of the cost savings generated by increased efficiencies passed on to consumers in terms of better services or reduced banking service fees are traded off with the expected costs implicit in the moral hazard and systemic risk arguments. In this paper we attempt to shed some light on this question through an empirical analysis that investigates the relationship between measures of the efficiency of a bank's operation on the one hand, and the size of the institution on the other.

More recently, regulatory features added by the Dodd–Frank Act (DFA)[1] introduced a variety of new policy levers, including capital surcharges, resolution plan requirements, consideration of systemic risk effects in mergers which specifically increased the emphasis on understanding of economies of scale and scope in large financial firms. That is, DFA requires the review of whether a proposed merger would lead to greater concentrated risks to financial stability. Regulators have encouraged researchers to better understand the social utility of the largest, most complex financial firms (Tarullo (2011)).

Some elaboration on what we mean by "too-big-to-fail" (TBTF) banks is also in order. During times of financial crisis banking supervisors have strong incentives to forestall the failure of large and highly interconnected financial firms due to the damage that such an event could pose to both the financial sector as well as the real economy. Unfortunately, as market participants anticipate that a particular firm may be protected in this way, this has the perverse yet highly rational effect of undermining market discipline and encouraging excessive risk-taking by the firm. Furthermore, it establishes economically unjustified incentives for a bank to become larger in order to reap this benefit. This results in a competitive advantage for such a large bank over its smaller competitors who may be perceived as lacking this implicit government safety net. Public sector bailouts are costly and politically unpopular and this issue has emerged as an enormous problem in the wake of the recent crisis. Therefore, as a tactical matter the state of the financial system has left supervisors with little choice but to use government resources to avoid failures of major financial institutions and accompanying destabilization of the financial sector. However, on a prospective basis supervisors have been directed to better address this issue through improved monitoring of systemically critical firms, with a view to preventing excessive risk-taking, and by strengthening the resilience of the financial system in order to minimize the consequences of a large firm being unwound.

A series of reforms have been proposed to address these problems. They include increasing capital requirements and limits upon leverage (e.g., Basel III), capping the size of banks, limiting the scope of banking activities, subjecting bank mergers and acquisitions to additional scrutiny, prescribing that banks draft "living wills" to plan their orderly unwinding, and requiring the federal government to proactively break up selected banks. These measures are not without their detractors, however. Feldman (2010), for example, casts doubt on the reforms focusing on size[2] by arguing even if such reforms could address TBTF, reforms that take aim at bank size directly might be bad policy because their costs could exceed their benefits. Moreover, the size of a bank may be positively related to other benefits. Large banks could offer cost advantages that would ultimately benefit society by taking advantage of scale economies in their service production processes. Wheelock and Wilson (2012), for example, concluded that most U.S. banks faced increasing returns to scale using their highly parameterized local linear estimator of banking services.

However, there may be problems with this perceived wisdom that large banks are large because of such scale economies for at least three reasons. First, some of the econometric work on economies of scale for banking, as in Hughes and Mester (1998), Hughes, Mester and Moon (2001), etc. find such benefits at all sizes of banks. Hughes and Mester (2008) summarize the extensive research findings in this regard. Second, we may simply not yet know very much about the presence of scale economies for today's unprecedentedly large banks. DeYoung (2010) emphasizes this point by arguing that the unique nature of today's large banks makes it difficult to apply statistical techniques to historical data to divine the extent of scale economies. It is clear that the financial sector has grown enormously in recent years. The question is "Why?" Banks indeed contribute to economic output through intermediation and have performed this economically useful function in many countries for hundreds of years, but value-added intermediation does not necessarily justify a large banking sector or banks whose current size is enormous by any historical standards. There are reasons to think that this sector may have become too big in the sense that too many of society's resources are allocated to it and may continue to contribute to a distortion in rents paid to those employed in the financial sector. Perceptions by creditors of banks that the government will protect them can lead the sector to grow inefficiently large as TBTF guarantees attract excessive funding to banks. These creditors understand that their bank investments are implicitly subsidized by the assurance of government bailouts should the bank begin to fail. For example, Tracey and Davies (2012) argues that there exists an "implicit funding subsidy" for TBTF banks.[3] Another point about the limits of our knowledge concerning the scale economies of large banks is that analysts face real challenges in measuring the "output" produced by banks. Since the banking sector provides loans, deposit and liquidity services it is a challenge to ensure that cross-firm comparisons are made controlling for these various service provisions, when economies of scale for the multi-output banking services technology is analyzed. Still another point is that the debate about TBTF and scale economies often presents the two in contradiction, when in fact they may complement one another. Some activities of a bank such may rely heavily on automation and thus may benefit from scale economies that enhance that bank's TBTF status.[4] The average cost of the large investments in these automated systems could be driven down by an increase in the volume of goods and services produced. Such automation-dependent products and services can generate a substantial portion portion of banking. Hence, greater scale activity could come with higher TBTF cost. The presence of economies of scale, from this perspective, suggests that policymakers sharpen their focus on fixing TBTF, (see Feldman (2010)).

The question of bank efficiency amongst the leading banking organizations in the U.S. is important as the banks must too comply with the stress test and capital plan requirements outlined by the Federal Reserve's Comprehensive Capital Analysis and Review (CCAR).[5] For estimating the impact of given stress testing scenarios, large banks have been relying on statistical models in order to quantify potential losses. The problem with this paradigm is that although it captures the social cost element it fails to capture the potential social benefits of bank scale and scope economies, as banks generally cannot incorporate these potential gains into their risk models. Our research contributes to a balanced analysis of this by considering efficiency measures.

Our paper analyzes the provision of banking services – the multi-output/multi-input technology that is utilized by banks in their role in the provision of banking services, including both balance-sheet financial intermediation businesses and off-balance-sheet activities. We focus on large banks, in particular the largest 50 financial institutions in the U.S. banking industry. The combined total assets of the largest 50 U.S. banks is close to 80 percent of the total assets of the U.S. banking system.[6] We examine the extent to which scale efficiencies exist in this subset of banks in part to address the issue of whether or not there are economic justifications for the notion that these banks may be "too-big-to-fail." Our empirical study is based on a newly developed dataset based on Call Reports from the FDIC for the period 1994–2013. We contribute to the post-financial crisis "too-big-to-fail" debate concerning whether or not governments should bail out large institutions under any circumstances, risking moral hazard, competitive imbalances and systemic risk. Restrictions on the size and scope of banks may mitigate these problems, but may do so at the cost of reducing banks' scale efficiencies and international competitiveness. Our study also utilizes a suite of econometric models and assesses the empirical results by looking at consensus among the findings from our various econometric treatments and models in order to provide a robust set of inferences on large scale banking performance and the extent to which scale economies have been exhausted by these large financial institutions. The analyses point to a number of conclusions. First, despite rapid growth over the last 20 years, the largest surviving banks in the U.S. have decreased in their level of efficiency. Second, we find no measurable returns to scale across our host of models and econometric treatments and in fact find negative correlation between bank size and the efficiency with which the banks take advantage of their scale of operations. In addition to the broad policy implications of our analysis, our paper also provides an array of econometric techniques, findings from which can be combined to provide a set of robust consensus-based conclusions that can be a valuable analytical tool for supervisors and others involved in the regulatory oversight of financial institutions.

The preceding section has provided a short discussion addressing previous studies related to our work. Section 2 describes the econometric models that will be estimated. In Section 3 we provide a description of our data set. A discussion of our empirical findings is presented in Section 4. Section 5 concludes.

2 Econometric models

In this section, we review our estimating framework. We will estimate second-order approximations in logs (translog) to a multi-output/multi-input distance function (see Caves, Christensen and Diewert (1982), Coelli and Perelman (1996)). The models we consider are linear in parameters. As our banking data constitute a balanced panel of banks and we are interested in a set of robust and consistent inferences from a wide variety of modeling approaches, we consider a number of different panel data estimators and assess the comparability of inferences from them. Our many treatments for various forms of unobserved heterogeneity can be motivated with the following classical model for a single output banking technology estimated with panel data assuming unobserved bank effects:

$$y_{it} = x_{it}\beta + \eta_i + u_{it} \qquad i = 1,\ldots,N;\ t = 1,\ldots,T \tag{5.1}$$

Here y_{it} is the response variable (e.g., some measure of bank output), η_i represents a bank specific Fixed Effect, x_{it} is a vector of exogenous variables and u_{it} is the error term.

In the classical Fixed Effects (FE) model for panel data, individual unobserved effects η_i are assumed to be correlated with the regressors x_{it}, while in the classical Random Effects (RE) model individual unobserved effects η_i are assumed to be uncorrelated with the regressors x_{it}. We also consider the Hausman and Taylor (1981) panel estimator. The H-T estimator distinguishes between regressors that are uncorrelated with the individual effects (x_{it}^1) and regressors that are correlated with the effects (x_{it}^2). As we have no time-invariant regressors in our study, the model becomes:

$$y_{it} = x_{it}^1\beta_1 + x_{it}^2\beta_2 + \eta_i + u_{it} \qquad i = 1,\ldots,N;\ t = 1,\ldots,T \tag{5.2}$$

We may interpret (5.1) or (5.2) as log-linear regressions, transformed from a Cobb-Douglas or translog function that is linear in parameters. In what follows, we do not distinguish between the *x's* that are, or are not, allowed to be correlated with the effects in order to reduce notational complexity. We do, however, make clear what these variables are in the empirical section. In order to move from a single to the multi-output technology considered in our empirical work we specify the multi-output distance function in the following way. Let the m outputs be $Y_{it} = \exp(y_{it})$ and the n inputs $X_{is} = \exp(x_{is})$. Then express the m-output, n-input deterministic distance function $D_O(Y,X)$ as a Young index, described in Balk (2008):

$$D_O(Y,X) = \frac{\prod_{j=1}^{m} Y_{it}^{\gamma_j}}{\prod_{k=1}^{n} X_{it}^{\delta_k}} \leq 1 \tag{5.3}$$

The output distance function $D_O(Y,X)$ is non-decreasing, homogeneous, and convex in Y and non-increasing and quasi-convex in X. After taking logs and rearranging terms we have:

$$-y_{1,it} = \eta_i + \sum_{j=2}^{m} \gamma_j y^*_{jit} + \sum_{k=1}^{n} \delta_k x_{kit} + u_{it}, i = 1,\ldots,N; \quad t = 1,\ldots,T \tag{5.4}$$

where $y^*_{jit,j=2,\ldots,m} = \ln(Y_{jit}/Y_{1it})$. After redefining a few variables, the distance function can be written as

$$y = X\beta + Z\eta + u \tag{5.5}$$

Here $y \in R^{NT}$ stacks the response variables across banks and time, the matrix $Z = I_N \otimes i_T \in R^{NT\times N}$ distributes the bank specific fixed effects (or the "incidence matrix" that identifies N distinct entities in a sample) that are stacked in the vector $\eta = (\eta_1, \eta_2, \ldots, \eta_N) \in R^N$, while $X = [x_{NT\times n}, y^*_{NT\times(m-1)}]$ contains both exogenous and endogenous variables and $U = (u_{it})^T \in R^{NT}$ is the stacked vector of error terms u_{it}.

However, the Cobb-Douglas specification of the distance function (Klein 1953) has been criticized for its assumption of separability of outputs and inputs and for incorrect curvature as the production possibility frontier is convex instead of concave. On the other hand, as pointed out by Coelli (2000), the Cobb-Douglas remains a reasonable and parsimonious first-order local approximation to the true function.[7] We also consider the translog output distance function, where the second-order terms allow for greater flexibility, proper local curvature, and lift the assumed separability of outputs and inputs. If the translog technology is applied, the distance function takes the form:

$$\begin{aligned}-y_{1it} = {}& \eta_i + \sum_{j=2}^{m} \gamma_j y^*_{jit} + \tfrac{1}{2}\sum_{j=2}^{m}\sum_{l=2}^{m} \gamma_{jl} y^*_{jit} y^*_{lit} + \sum_{k=1}^{n} \delta_k x_{kit} + \tfrac{1}{2}\sum_{k=1}^{n}\sum_{p=1}^{n} \delta_{kp} x_{kit} x_{pit} \\ & + \sum_{j=2}^{m}\sum_{k=1}^{n} \theta_{jk} y^*_{jit} x_{kit} + u_{it}, \qquad i = 1,\ldots,N; t = 1,\ldots,T\end{aligned} \tag{5.6}$$

This can be written in the form of Equation (5.1). Here X contains the cross-product terms as well as the own n input m-1 normalized output terms. $X = [x_{NT\times n}, y^*_{NT\times(m-1)}, xx_{NT\times(n\times(n+1)/2)}, y^*y^*_{NT\times((m-1)\times m/2)}, xy^*_{NT\times(m-1)\times n)}]$, the latter of which appear in their normalized form owing to the homogeneity of the output distance function.

In the translog specification, our focus should be on the following key derivatives, which correspond to the input and output elasticities. The derivatives are expressed as follows in Equations (5.7) and (5.8).

$$s_p = \delta_p + \sum_{k=1}^{n} \delta_{kp} x_k + \sum_{j=2}^{m} \theta_{pj} y_j^*, \quad p = 1,2,\ldots,n \tag{5.7}$$

$$r_j = \gamma_j + \sum_{l=2}^{m} \gamma_{jl} y_j^* + \sum_{k=1}^{n} \theta_{kj} x_k, \quad j = 2,\ldots,m \tag{5.8}$$

2.1 Frontier estimation methodology

In this subsection, we describe our estimation methodology utilizing the semiparametric efficiency estimators summarized in Sickles (2005). We utilize Equation (5.2) and consider cases in which u and (η, x_1, x_2) are independent but there is a level of dependency among the effects and the regressors. Equation (5.1) can be reinterpreted as a stochastic panel production frontier model introduced by Pitt and Lee (1981) and Schmidt and Sickles (1984). Although we may be on somewhat solid footing by invoking a central limit argument to justify a Gaussian assumption on the disturbance term u_{it}, we may be far less justified in making specific parametric assumptions concerning the distribution of the η_i term, which in the stochastic frontier efficiency literature is interpreted as a normalized radial shortfall in a bank's performance relative to the best-practice performance it could feasibly attain. While we can be confident in restricting the class of distributions of the inefficiency term to those that are one-sided (see the inequality in Equation (5.3)), the heterogeneity terms are intrinsically latent and unobservable components and we encounter problems regarding identifiability of these parameters (Ritter and Simar (1997)). The additional model we use in our analyses is a semiparametric efficient (SPE) estimator and is well-suited to provide us with robust point estimates and minimum standard errors when we are unwilling to use parametric assumptions for the distribution of the heterogeneity terms and their dependency with either all or some of the regressors. The general approaches to deriving such semiparametric efficient estimators is discussed at length in Newey (1990) and Pagan and Ullah (1999), as well as in a series of papers by Park, Sickles and Simar (1998, 2003, 2007). Interested readers can find the derivations for the SPE panel stochastic frontier estimators we utilize in our empirical work below in the cited papers. The framework for deriving all of the estimators is somewhat straightforward and has much in common across the different stochastic assumptions on which the different SPE estimators are based.

We utilize a particular SPE estimator in our analyses. This estimator is detailed in Park et al. (1998). We refer to this as the PSS1 estimator and it is an extension of the estimator introduced in Park and Simar (1994), which assumed the effects to be independent of all of the regressors. We assume in the specification (5.2) that the set of regressors $x_{1,it}$ is conditionally independent of the individual unobserved random effects η_i given the set of correlated regressors $x_{2,it}$:

$$f(\eta, x_1, x_2) = h(\eta, x_2) g(x_1 | x_2) \tag{5.9}$$

Furthermore, it is assumed that η_i depends on $x_{2,it}$ only through its long-run movement:

$$\mathrm{h}(\eta_i,\ \mathrm{x}_{2,it}) = \mathrm{h}_M(\eta_i, \bar{x}_{2,it}) p(x_{2,it}) \tag{5.10}$$

Here $\mathrm{h}_M(\eta_i, \bar{x}_{2,it})$ is a nonparametric multivariate density specified using kernel smoothers. We will discuss our strategy for selection of the variables that are portioned into $x_{1,it}$ and $x_{2,it}$.

In addition to the PSS1 SPE estimator, we consider an alternative approach that allows for time-varying heterogeneity, interpreted in the stochastic frontier literature as a normalized level of technical efficiency. The approach is parametric. Battese and Coelli (1992), henceforth BC, consider a panel stochastic frontier production function with an exponential specification of time-varying firm effects:

$$\begin{aligned} Y_{it} &= f(X_{it}, \beta) \exp(u_{it} - \eta_{it}) \\ \eta_{it} &= \{\exp[-\varsigma(t-T)]\} \eta_i \end{aligned} \tag{5.11}$$

where $u_{it} \sim NID(0, \sigma_u^2)$ and $\eta_i \sim NID^+(0, \sigma_v^2)$ are normal i.i.d. and non-negative truncated normal i.i.d., respectively. Maximum likelihood estimators of the model parameters can be derived and mean technical efficiency can be constructed.[8]

2.2 Quantile regression

A final class of estimator we consider in our empirical analyses of banking performance is the panel quantile regression model. The τ^{th} conditional quantile function of the response y_{it}, the analog to Equation (5.1), can be written as:

$$Q_y(\tau | Z, X) = X\beta(\tau) + Z\eta + u \tag{5.12}$$

Note that in the model, the effects $\beta(\tau)$ of the covariates X are allowed to depend upon the quantile τ. The vector η is intended to capture individual specific sources of unobserved heterogeneity that are not adequately controlled for by other covariates. The estimates of the individual specific effects (η's) are restricted to be invariant with respect to the quantile but are allowed to be correlated with the x's as they are modeled as fixed effects. As pointed out in Galvao (2011), in settings in which the time series dimension is relatively large allowing quantile specific fixed effects is not feasible.

Koenker (1984) considered the case in which only the intercept parameter was permitted to depend upon the quantile and the slope parameters were constrained to be identical over selected quantiles. The slope parameters are estimated as regression L-statistics and the individual effects are estimated as discretely weighted L-statistics.

The model we apply in this paper is the quantile regression fixed effects model for panel data developed in Koenker (2004), which solves the following convex minimization problem:

$$(\widehat{\beta},\widehat{\eta})^T = \underset{\beta,\eta}{\arg\min}\left\{\sum_{k=1}^{K}\sum_{i=1}^{N}\sum_{t=1}^{T} v_k\rho_\tau(y_{it} - x_{it}\beta(\tau_k) - \eta_i z_{it})\right\} \tag{5.13}$$

where k indexes the K quantiles $\{\tau_1, \tau_2, \ldots, \tau_k\}$, $\rho_\tau(u) \equiv u(\tau - I_{u<0})$ is a piecewise linear quantile loss function as defined in Koenker and Bassett Jr (1978), and v_k are weights that control the influence of the quantiles on the parameter estimates. The choice of the latter are analogous to discreetly weighed L-statistics (Mosteller 1946), a common choice of which is Tukey's trimean (Koenker 1984).

3 Data

The bank sample is from the top 50 U.S. banks by total book value of assets (TBVA), as of the third quarter of the year 2013, from quarterly Call Reports. More precisely, we have quarterly data from 1Q1994 to 3Q2013, obtained from the "Consolidated Reports of Condition and Income for a Bank with Domestic and Foreign Offices – FFIEC 031" regulatory reports, expressed on a pro-forma basis that go back in time to account for mergers. In order to illustrate, if a bank in 2008 is the result of a merger in 2008, pre-2008 data is merged on a pro-forma basis (i.e., the other non-surviving bank's data will be represented as part of the surviving bank going back in time). The rationale behind this methodology is to create a long historical data set that controls for survival bias, and also that does not exhibit a distorted measure of banks' growth. U.S. bank regulators use this data in order to estimate risk measurement models, such as the Bank Capital-at-Risk Model (Frye and Peltz 2008), which is the basis of risk dashboards used for centralized bank supervision. While this sample design is not a common practice amongst academics, this does reflect methodologies used by banks in calibrating credit risk models, such as those used for Basel III and for CCAR.[9]

Although we intended to analyze the top 50 U.S. commercial banks, due to missing and questionable data entry, we ended up using 44 of these banks in our analyses. The five output and six input variables used to estimate the distance function using both stochastic frontier analysis and quantile regression are:

Real Estate Loans (REL)
Commercial and Industrial Loans (CIL)

Consumer Loans (CL)
Securities (SC)
Off-Balance-Sheet Activities (OFF)
Premises & Fixed Assets (PFA)
Number of Employees (NOE)
Purchased Funds (PF)
Savings Accounts (SA)
Certificates of Deposit (CD)
Demand Deposits (DD).

The risk proxies are:

CREDIT RISK: Gross Charge-off Ratio (CR)
LIQUIDITY RISK: Liquidity Ratio (LR)
MARKET RISK: Trading Revenue Deviation to Trading Book Ratio (MR)

Before further providing the descriptive statistics on our variables, we would like to draw attention to our contribution to the banking efficiency in terms of a control variable, i.e., Market Risk (MR) proxy, which we have used in our analyses. Market risk results from holding or taking positions in interest rates, foreign exchange, equities, commodities, and credit spreads. While the core function of traditional banking is to accept deposits and make loans, large banks also take market risk on their trading books and make trading revenues. Loosely speaking, the banking book comprises lending activities, whereas the trading book comprises trading securities, over-the-counter (OTC) derivatives[10] and market-making activities. Notwithstanding the fact that, the 2007–2008 financial crisis was initiated by a U.S. housing crisis, OTC derivatives which are mainly reported on banks' trading books contributed to amplifying various problems and provided channels for systemic risk to propagate (Gregory 2014, 3). The key differences between the trading and banking books relate to holding intent, liquidity and mark-to-market valuation. It has been evidenced that traditional banking business of accepting deposits and making loans has declined significantly in the U.S. (Allen and Santomero 2001). The evidence continues to prevail in the ratio of the size of the trading book to total loans (i.e., traditional lending business) for top U.S. banks even after the 2007–2008 financial crisis (Figure 5.1).

Regulatory capital requirements for the banking and trading books differ significantly. As trading book positions are daily marked-to-market and actively hedged by the banks, they are not intended to be held for an extended period of time. Hence, the regulatory capital charges for such positions have been based on the price volatility. The first market risk regulatory capital requirements to recognize this fact were introduced in 1996 (Basel I Amendment). The 1996 amendment required banks to estimate a risk measure – the so-called

Value-at-Risk (VaR) – for trading book positions over a ten-day time horizon. However, during the 2007–2008 financial crisis, losses in many banks' trading books have been significantly higher than the minimum capital requirements under the market risk rules (BCBS 2009a). Across global banks, trading book losses totaled over $900 billion over 2007–2009 (Haldane 2009). The explanation was straightforward: when markets remained liquid and asset prices rose, banks gained from mark-to-market trading book valuations, but when asset prices fell during a financial crisis, market-maker banks incurred billion-dollar losses on their trading books. This was clearly the case for major U.S. banks. Before the crisis, the top five U.S. banks rarely reported quarterly trading losses but incurred multiple billion-dollar losses during the crisis quarters (Figure 5.2).

In response to the financial crisis, the Basel Committee on Banking Supervision (BCBS) introduced incremental changes to the current VaR based trading book framework in 2009 (also known as Basel 2.5, BCBS 2009a).[11] The short-term fix was to recognize the credit risk in the trading book with an incremental risk capital charge (IRC) for unsecuritized credit products, and a comprehensive risk measure (CRM) for tranched credit products. Additionally, BCBS required banks to calculate a stressed VaR, taking into account a one-year observation period relating to significant losses, which must be calculated in addition to the Value-at-Risk based on the most recent one-year observation period. The additional stressed value-at-risk requirement was incorporated to help reduce the procyclicality of the minimum capital requirements for market risk.

While these additional measures were meant to capture the real risk exposures of trading books, BCBS had agreed that the additional measures were not sufficient and planned to carry out a more fundamental review of the market risk framework, including the use of VaR estimates as the basis for the minimum capital requirement. The initial proposal[12] was released in May 2012 (BCBS 2012) and focused on key areas: such as, the trading book/banking book boundary, expected shortfall (ES) measure as an alternative to VaR, and a comprehensive incorporation of the risk of market illiquidity among other things. The importance of incorporating the risk of market illiquidity is a key consideration in banks' regulatory capital requirements for trading portfolios. Before the introduction of the Basel 2.5 changes, the entire market risk framework was based on an assumption that trading book risk positions were liquid, i.e., that banks could liquidate these positions over a ten-day horizon. The recent crisis proved this assumption to be false. That is, during the financial crisis, banks experienced significant illiquidity in a wide range of credit products held in the trading book: hence, they were forced to retain exposures for prolonged periods of time.

Having stated the problems encountered for banks' trading portfolios during the crisis and the recent regulatory responses to the trading book related issues, we included a market risk proxy in our efficiency models in order to recognize the risk exposure of banks' trading books. To the best of our knowledge, this is the first paper which uses the variability of *unexpected* trading revenue as the market risk proxy in the banking efficiency literature. Ideally, one should use the quarterly Value-at-Risk (VaR),[13] which is an average of daily reported VaR's in a given quarter to proxy a bank's market risk exposure; however, as daily VaRs are not available to us, we follow Jorion (2002), who demonstrated that a bank's expected absolute value of "*unexpected* trading revenue" is proportional to the dispersion of Value-at-Risk (VaR) if the trading revenue is distributed symmetrically around zero. Following Jorion, we remove an estimate of the mean of the trading revenue (i.e., moving average of the last four quarters) in order to calculate the variability of trading revenue, which is proxied as the absolute value of unexpected trading revenue. We then divide the absolute value of unexpected trading revenue by the gross sum of trading assets and trading liabilities to calculate the market risk proxy. That is,

$$\mathrm{MR} = \frac{\left|\begin{array}{c}\text{Deviation from the moving average of last}\\ \text{4 quarters of trading revenue}\end{array}\right|}{(\text{Trading Assets+Trading Liabilities})}$$

Returning to descriptive statistics, Table 5.1 summarizes key variables as of 3Q2013, from the Call Reports for the top nationally chartered banks in the U.S. by *total book* value of assets (TBVA) at this time. We display details on the top ten out of 50 by TBVA in descending order (JP Morgan Chase, Bank of America, Wells Fargo, Citigroup, US Bank, Capital One, Bank of New York Mellon, PNC, State Street and HSBC) and distributional statistics on the top 50. The data is extremely skewed in terms of size as measured by TBVA, with the average of the top four in TBVA each in excess of the 95th percentile of $1.45 trillion, and the top ten comprising $8.14 trillion (or 79.27 percent) out of the 10.27$ trillion total, as compared to the median TBVA of $81.35 ($233.47) billion. There is similar extreme skew by the value of Total Loans (TL), with the average of the top four in TBVA in excess of the $751.44 billion 95th percentile of TL, and the top ten comprising $3.81 trillion or (74.08 percent) out of the $5.14 trillion total, as compared to a median TL of $40.64 ($116.80) billion. We observe more extreme skew than even TBVA in the value of trading revenue deviation, with the average of the top four in significant excess of the $242.05 million 95th percentile of trading revenue deviation, and the top ten comprising $1.84 billion or (90.12 percent) out of the $2.04 billion total, as compared to a median trading revenue deviation of $46.33 million ($3.16 million). Similarly, total gross charge-offs are skewed toward the largest banks, with the average of the top four in TBVA each in excess of the $1.81 billion 95th percentile of gross charge-offs, and the top ten comprising $9.93 billion (or 83.99 percent) out of the $11.82 billion total, as compared to median gross charge-offs of $48.74 million ($268.69 million). Finally, for the dollar measures, total cash balances are very much concentrated in the largest banks, with the average of the top four in TBVA in excess of the $244.53 billion 95th percentile of total cash balances, and the top ten comprising $1.66 trillion (or 87.53 percent) out of the $1.89 trillion total, as compared to median total cash balances of $5.45 ($42.99) billion. Gross Charge-off ratios (CR) for many of the top ten are on the high side relative to the center of the distribution, eight of them above (ranging in 0.11–0.67 percent) the median in the broader sample of 0.12 percent (0.16 percent). There is a similar pattern with respect to liquidity ratios (LR), with many of the top ten on the high side relative to the center of the distribution, eight of them above (ranging in 10.19–49.73 percent) the median in the broader sample of 7.41–13.04 percent. Figures 5.3 through 5.7 represent several of these measures in time series on from the first quarter of 1994 until the third quarter of 2013.

Figure 5.1 shows the ratio of the trading book to total loans across the U.S. top 44 out of 50 banks from 1994Q1 to 2013Q3. This ratio fluctuates from five percent to eight percent in the 1990s, and sharply surge up to 15 percent in early 2000s. It reaches the peak of around 25 percent in 2007 and drops to 17 percent in less than two years. The ratio continues decreasing in most recent years. Figure 5.2 displays the trading revenue trend for the top five banks. These banks show similar fluctuations in time trend though some banks have greater variations than others do. These banks rarely experience negative trading revenues but incurred significant amount of dollar losses during the crisis quarters. Figure 5.3 shows the TBVA across the U.S. 44 out of 50 largest banks over time, reflecting the growth in the banking industry overall as well as of the largest banks, with TBVA increasing smoothly from around just under $4 trillion in the early 1990s, to about $10 trillion during the recent financial crisis and declining about one trillion until 2010 and then bouncing over $10 trillion recently. Figure 5.4 shows the quarterly TL from over this period, which shows a similar trend to TBVA, a secular upward trend of growth (from about $2.5 to nearly $5.5 trillion in 2008), as well the financial crisis, reflected dips of about $1 trillion in the period 2008 to 2009, and increased slowly since then. In Figure 5.5, the time series of CRs clearly reflects the credit cycle, with previous peaks of 0.4 percent around early 2000s, and alarmingly near 1 percent by the end of 2009. On the other hand, in Figure 5.6, LRs display a markedly different pattern over time as compared to CRs, a secular decline from around ten percent, at the beginning of the sample period, to around 6 percent from 1997 to 2001, and reaching up to about 16 percent after 2007 and fluctuating since then to 17 percent at the end of the sample period. Finally, in Figure 5.7 we see the ratio of deviation of trading revenue from the moving average of the previous four quarters to the trading book, displaying yet another different pattern to the other risk measures: it shows one mode around year 2000 and another peak in the year 2007 and sharp decline since then. In Figures 5.8 through 5.12 we show the

Table 5.1 Characteristics of top 50 banks by total book value of assets as of Q32013 (Call Report Data 1994Q1–2013Q3)

	Bank	Book Value of Assets	Banking Book Loans	Gross Charge-offs	Cash Balances	Trading Revenue Deviation	Trading Book	Charge-off Ratio	Liquidity Ratio	Trading Revenue Deviation Ratio
Details on top 10	J.P. Morgan Chase	2,122,287,068	725,503,268	1,642,129	642,077,208	1,120,000	371,149,000	0.23%	30.25%	0.30%
	Bank of America	1,621,455,000	925,878,000	2,205,000	165,301,000	298,395	113,693,000	0.24%	10.19%	0.26%
	Wells Fargo & Co.	1,380,697,455	811,970,656	1,323,356	211,042,671	128,750	48,215,000	0.16%	15.29%	0.27%
	Citigroup	1,346,413,607	603,523,317	2,594,637	322,674,064	217,909	186,016,138	0.43%	23.97%	0.12%
	U.S. Bancorp	356,590,456	233,535,765	450,397	15,085,200	19,259	1,319,346	0.19%	4.23%	1.46%
	Capital One Financial Corp	313,154,981	192,463,829	1,296,423	30,757,816	12,205	790,465	0.67%	9.82%	1.54%
	Bank of New York Mellon	309,488,944	38,396,960	2,457	153,895,620	5,000	13,514,000	0.01%	49.73%	0.04%
	PNC Financial	298,485,621	195,566,120	340,744	15,712,887	19,478	4,643,815	0.17%	5.26%	0.42%
	State Street	212,689,010	15,636,947	1	47,486,430	12,903	10,728,408	0.00%	22.33%	0.12%
	HSBC	181,762,250	64,552,314	73,301	51,422,421	3,159	37,506,127	0.11%	28.29%	0.01%
Statistics on top 50	Minimum	19,301,507	9,305,195	1	283,329	0	1	0.00%	1.38%	0.00%
	5th Percentile	20,889,429	10,523,202	1	766,956	5	1,401	0.00%	1.63%	0.01%
	25th Percentile	25,174,400	15,707,171	12,304	1,756,586	775	114,623	0.08%	4.29%	0.20%
	Median	81,348,361	40,637,447	48,741	5,446,157	3,160	577,647	0.12%	7.41%	0.41%
	Average	233,465,451	116,801,277	268,668	42,985,668	46,327	18,400,566	0.16%	13.04%	1.56%
	75th Percentile	171,570,765	74,088,496	148,414	20,443,619	13,257	1,934,628	0.18%	16.77%	1.08%
	95th Percentile	1,452,924,719	751,443,484	1,810,990	244,532,089	242,054	135,389,941	0.50%	47.13%	10.93%
	Maximum	2,122,287,068	925,878,000	2,594,637	642,077,208	1,120,000	371,149,000	0.76%	49.73%	17.67%
	Standard Deviation	461,032,705	217,045,601	592,341	112,345,609	175,235	63,699,072	0.16%	12.92%	3.52%
	Skewness	10.4004	9.1999	9.5389	20.3433	33.6645	23.6458	8.7609	4.4204	14.4414
	Kurtosis	2.9058	2.7271	2.7461	4.0299	5.4930	4.4868	2.2786	1.5191	3.4713
	Grand Total	8,143,024,392	3,807,027,176	9,928,445	1,655,455,317	1,837,056	787,575,299	0.26%	20.33%	0.23%

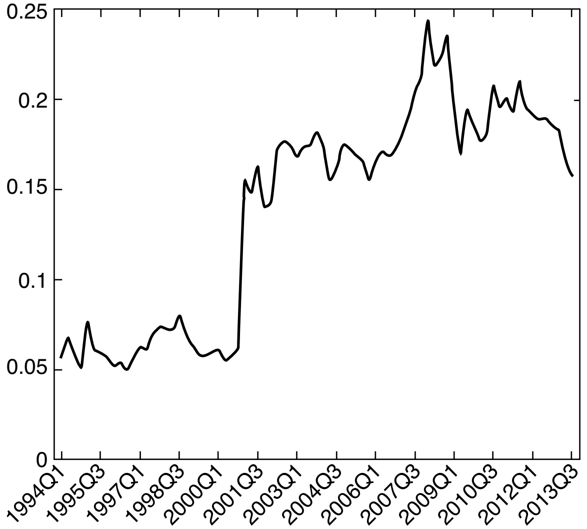


Figure 5.1 Ratio of trading book to total loans as of 2013Q3

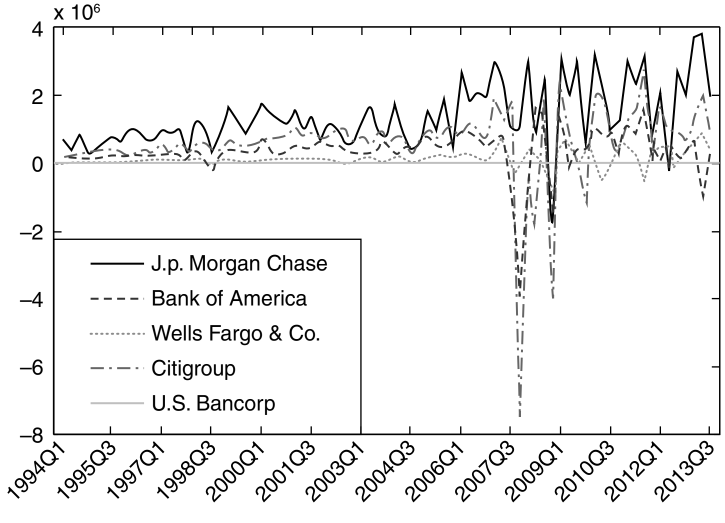


Figure 5.2 Trading book revenue for top five US banks as of 2013Q3 (in thousand $)

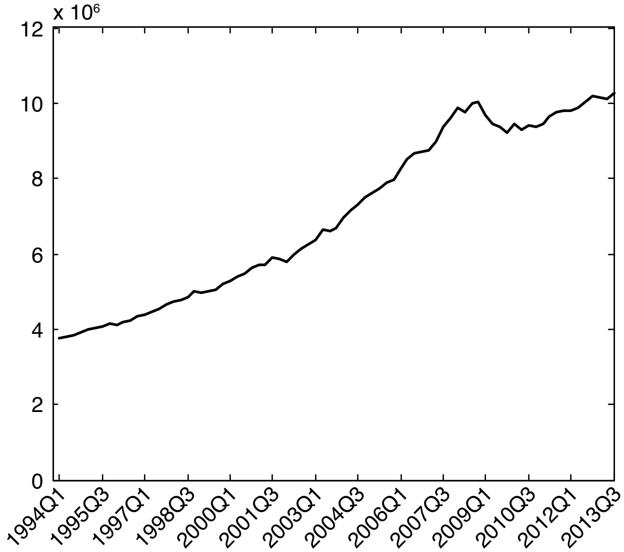


Figure 5.3 Total book value of assets (in million $)

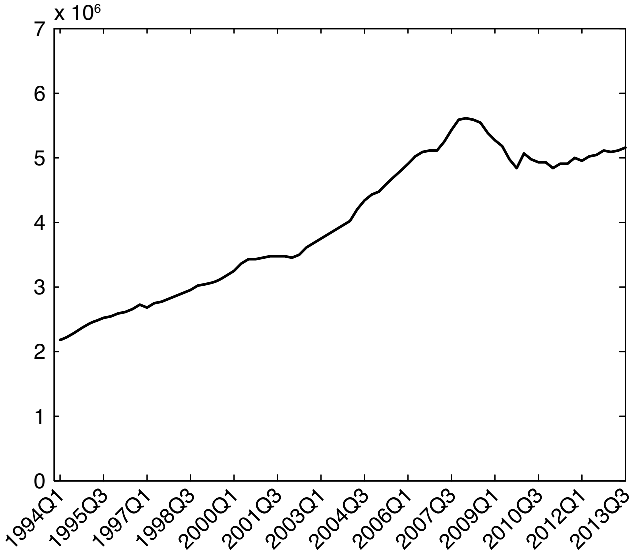


Figure 5.4 Total loans (in million $)

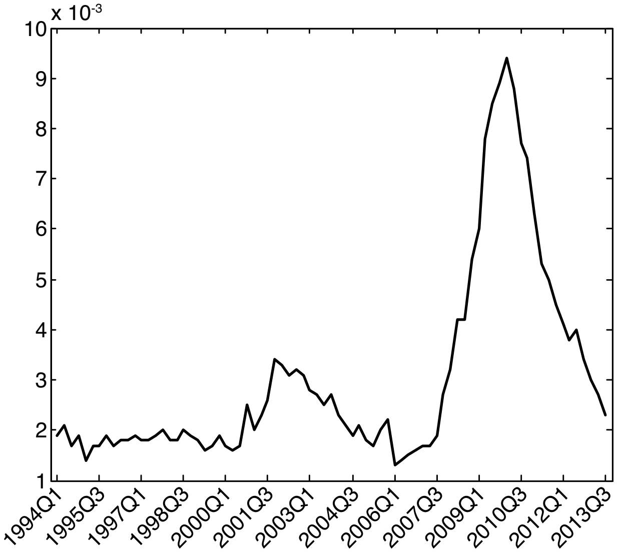


Figure 5.5 Average ratio of total charge-off to total loans

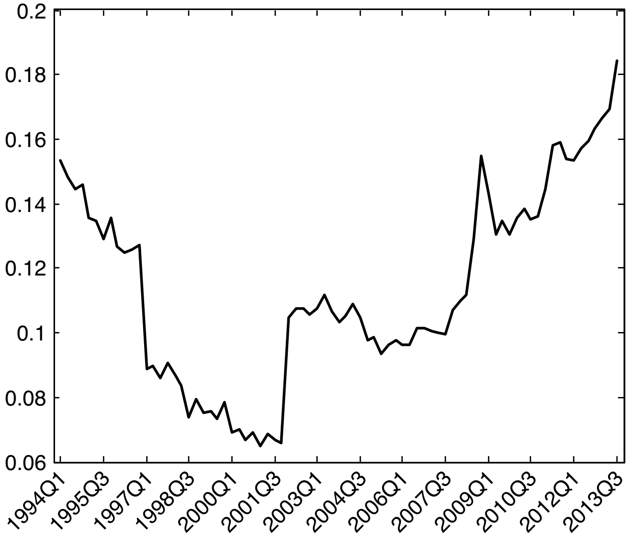


Figure 5.6 Average liquidity ratios

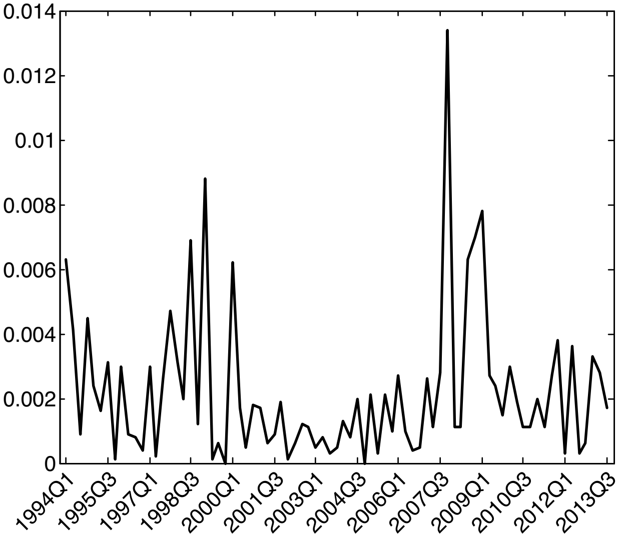


Figure 5.7 Average trading revenue deviation to trading book

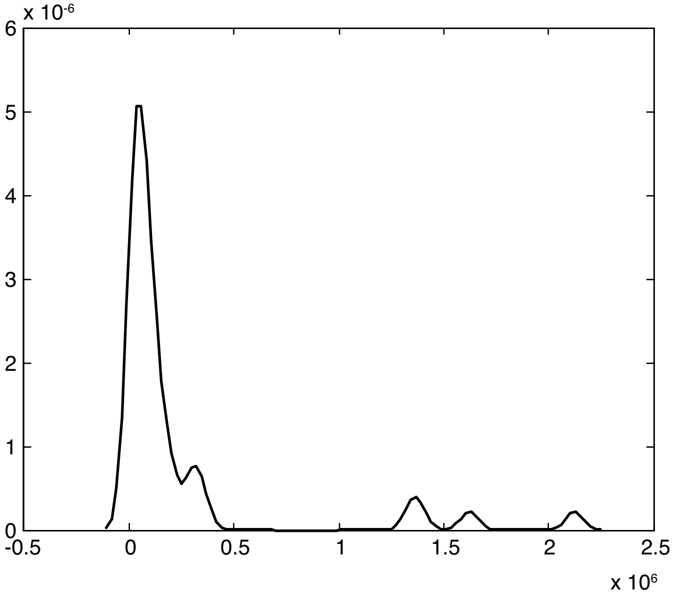


Figure 5.8 Distribution of total book value of assets as of 2013Q3

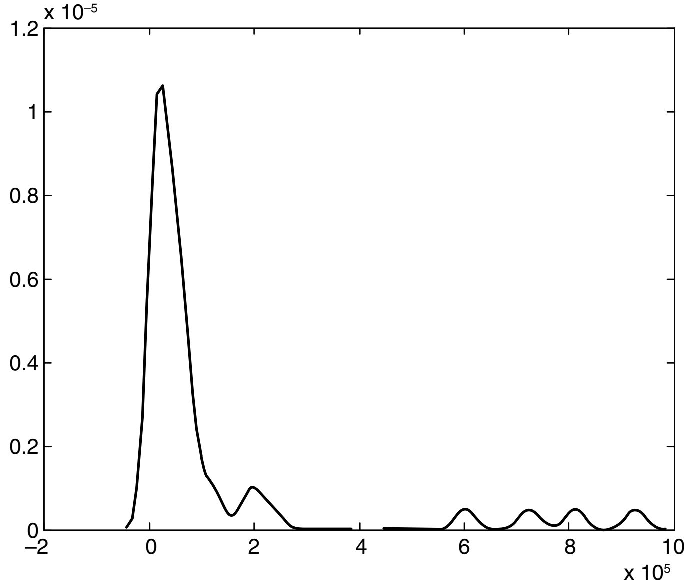


Figure 5.9 Distribution of total loans as of 2013Q3

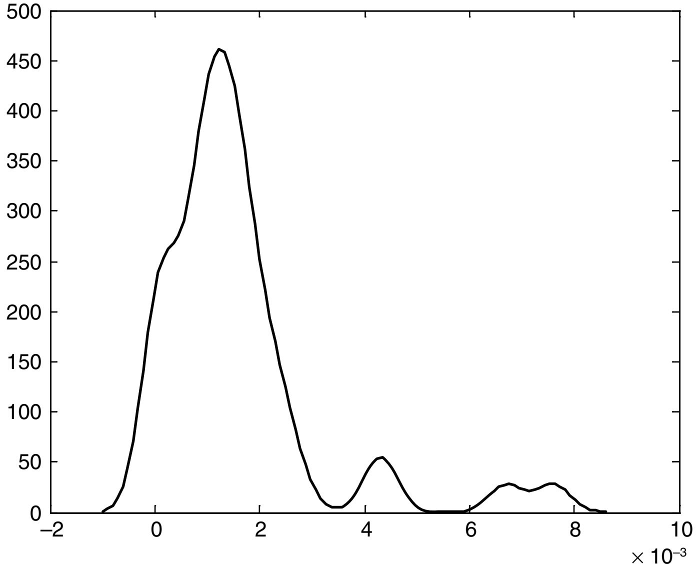


Figure 5.10 Distribution of total charge-off to total loan ratios as of 2013Q3

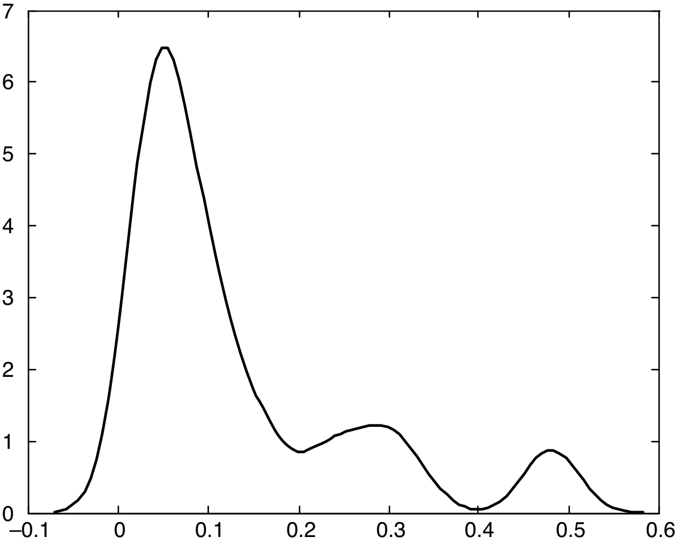


Figure 5.11 Distribution of liquidity ratios as of 2013Q3

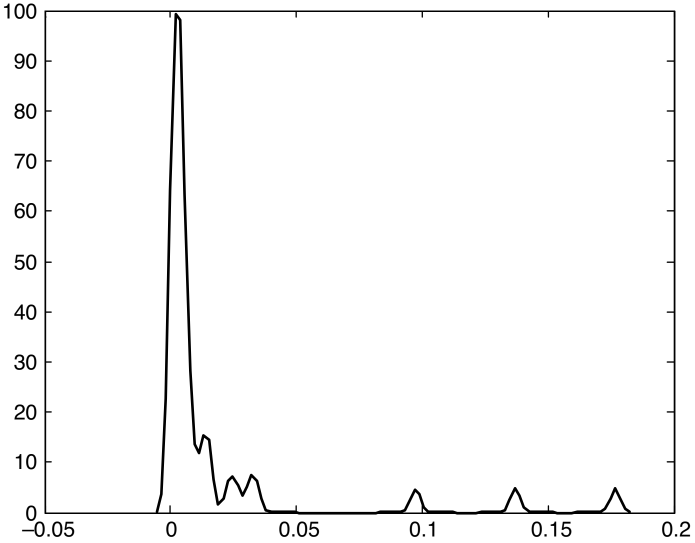


Figure 5.12 Distribution of trading revenue deviation to trading book as of 2013Q3

distribution of the five measures, analyzed in Table 5.1 across the largest banks as of 3Q2013. The right skewness in all of these variables is evident.

4 Estimation results

Our specifications of the translog output distance functions are based on the intermediation interpretation of banking services wherein banks utilize deposits and other input factors to provide loan services as their outputs, (see Sealey and Lindley (1977)). The alternative production approach views deposits as outputs as opposed to inputs proposed by Baltensperger (1980).

Anticipating the discussion to follow, the overall conclusion of our empirical analyses is that the largest surviving banks – in spite of tremendous growth in the last 20 years – have experienced a diminished capacity to provide loan services as they took on increasing levels of risk. This is reflected in a decline in efficiency as implied by the econometric models that allow efficiency levels to vary temporally. In addition, larger banks have lower scale efficiency levels. There is no evidence of scope economies. Finally, there is no evidence of economies of scale for the large banks in our sample.

The elasticities of six inputs and three outputs are evaluated at the sample mean of the data points, in Table 5.2, where the standard errors are reported in parentheses. We utilize a nonparametric bootstrap following Efron and Tibshirani (1986) , which is implemented through 1,000 iterations where in each run, 44 banks are chosen with replacement and 79 quarters are chosen with replacement, and the model is re-estimated. Since our dataset is mean deflated prior to estimating the distance function, the first derivatives expressed in Equations (5.7) and (5.8) will simply be equal to the first-order coefficients when evaluated at the sample mean.

The elasticity estimates shown in Table 5.2 are consistent with the monotonicity assumption. The six inputs' elasticities have negative signs, and the three outputs' elasticities have positive signs. Alternatively, all of the input variables (Premises and Fixed Assets, Number of Employees, Purchased Funds, Savings Accounts, Certificates of Deposit and Demand Deposits) contribute positively to the output, albeit varying in magnitude. Compared with the other inputs, SA and DD have the greatest impact. NOE is also an important input source albeit it has less impact than SA and DD; while the estimates of PFA and CD are similar in magnitude. PF has the smallest impact in all the inputs.

Across most models, our estimates suggest no evidence of increasing returns to scale since the numbers vary closely around one.

Turning our attention to the controls for risk, which are displayed in the last three rows of Tables 5.4 and 5.5 in the appendix, we observe that in all have generally positive signs on coefficient estimates, which have the interpretation

Table 5.2 The elasticity estimates evaluated at sample mean

	FE	RE	FEIV	REIV	H-T	PSS1	BC	QR(50%)
PFA	–0.0486	–0.0519	–0.0903	–0.0875	–0.0500	–0.0437	–0.0253	–0.0640
	(0.0501)	(0.0506)	(0.0490)	(0.0809)	(0.0485)	(0.0433)	(0.0562)	(0.0405)
NOE	–0.1745	–0.2121	–0.0978	–0.1571	–0.1839	–0.1601	–0.2116	–0.1204
	(0.0637)	(0.0595)	(0.0614)	(0.0728)	(0.0650)	(0.0858)	(0.0749)	(0.0603)
PF	–0.0215	–0.0202	–0.0224	–0.0206	–0.0212	–0.0224	–0.0141	–0.0195
	(0.0039)	(0.0030)	(0.0049)	(0.0049)	(0.0051)	(0.0051)	(0.0033)	(0.0029)
SA	–0.5519	–0.5582	–0.5453	–0.5529	–0.5532	–0.5611	–0.5526	–0.5905
	(0.0400)	(0.0489)	(0.0401)	(0.0576)	(0.0440)	(0.0376)	(0.0644)	(0.0415)
CD	–0.0586	–0.0569	–0.0443	–0.0423	–0.0583	–0.0606	–0.0712	–0.0778
	(0.0135)	(0.0136)	(0.0119)	(0.0170)	(0.0127)	(0.0132)	(0.0192)	(0.0120)
DD	–0.0828	–0.0984	–0.0894	–0.1197	–0.0861	–0.0971	–0.1553	–0.1086
	(0.0286)	(0.0363)	(0.0327)	(0.0487)	(0.0314)	(0.0277)	(0.0468)	(0.0242)
REL	0.4028	0.3793	0.4247	0.3878	0.3982	0.4125	0.3029	0.4823
	(0.0495)	(0.0480)	(0.0548)	(0.0348)	(0.0595)	(0.0635)	(0.0606)	(0.0348)
CIL	0.2105	0.2172	0.2254	0.2283	0.2117	0.2139	0.2266	0.1810
	(0.0400)	(0.0366)	(0.0471)	(0.0333)	(0.0360)	(0.0446)	(0.0228)	(0.0287)
CL	0.0817	0.0814	0.0495	0.0581	0.0819	0.0737	0.0707	0.0719
	(0.0253)	(0.0186)	(0.0206)	(0.0208)	(0.0237)	(0.0303)	(0.0183)	(0.0217)
SC	0.2604	0.2700	0.2704	0.2829	0.2622	0.2678	0.3242	0.2462
	(0.0270)	(0.0307)	(0.0309)	(0.0201)	(0.0291)	(0.0366)	(0.0497)	(0.0220)
OFF	0.0446	0.0521	0.0299	0.0428	0.0461	0.0320	0.0757	0.0185
	(0.0140)	(0.0203)	(0.0109)	(0.0128)	(0.0122)	(0.0125)	(0.0240)	(0.0103)
RST	0.9379	0.9978	0.8895	0.9801	0.9527	0.9451	1.0301	0.9809
	(0.0661)	(0.0281)	(0.0826)	(0.0400)	(0.0545)	(0.0690)	(0.0247)	(0.0457)

that all else equal, risk-taking activities decrease output, as more risk is detrimental and reduces the capacity of the banks to make loans. The magnitudes of the coefficient estimates of Credit Risk (CR) are around ten times smaller than Liquidity Risk (LR). As LR is proxied by the liquidity ratio (cash balance/total assets) one might first expect a negative sign on the coefficient since the positive signs indicated by all of the estimators indicate that increases in LR reduce the level of intermediation services provided by the bank. It is clear from our estimates that these banks are not managing their liquidity optimally, controlling for market and credit risk. The positive sign for coefficient estimates of Market Risk (MR) suggest that as banks move from traditional banking (i.e., lending business) to trading book activities, banks have become less efficient in lending.

Coefficient estimates on all of the three risk proxies are generally the same across models using both stochastic frontier analysis and quantile regression. The positive signs on the coefficient estimates indicate that greater LR or

MR inhibits output. The estimates on MR are generally much less substantial across models; the estimates on LR consistently have more substantial across models than the other two risk proxies. These results regarding LR and MR support the policy argument that banks should be restricted from engaging in highly risky activities, such as proprietary trading, and encouraged to maintain an appropriate liquidity ratio. More generally, our results, taken in totality, lead to the sensible implication that banks which stray from their core competencies will provide less intermediation services and should shrink over time.

In Figure 5.13 and Table 5.5, we summarize the estimation results of the quantile regression Fixed Effects model for panel data. We estimate these models in the R statistical programming language (R Team 2010) using the quantreg package by Koenker (2009), which the authors adapt and extend in order to produce longitudinal data results, as well as to produce more reliable statistical inference. From Figure 5.13, we can see that the quantile regression estimates on the elasticities, represented in black lines, are compatible with those from the Fixed Effects model, which are denoted in the red lines. The elasticity estimates do not vary significantly across quantiles, but the estimates on Credit Risks and Liquidity Risks have displayed a distinctive increasing pattern.[14]

Economies of scope, displayed in Table 5.3, are constructed following Hajargasht, Coelli and Rao (2008), who derive the expression for economies of scope in terms of the derivative of the distance functions, utilizing the duality between the cost and input distance functions. The economies of scope between outputs i and j can be calculated using the derivatives of the output distance function as follows:

$$C_{yy}/C = D_yD'_y - D_{yy} + D_{yx}[D_{xx} + D_xD'_x]^{-1}D_{xy} \tag{5.14}$$

Our dataset is centered on the geometric mean of all observations. Results are essentially the same when we center at the median time period as well. This enables us to more transparently interpret the translog results. Economies of scope evaluated at the sample geometric means for the median time period can be calculated following this formula in Equation (5.15). A positive sign represents scope diseconomies.

$$D_yD'_y - D_{yy} + D_{yx}[D_{xx} + D_xD'_x]^{-1}D_{xy} = \begin{bmatrix} \gamma_1 - \gamma_{11} & \cdots & -\gamma_{1m} \\ \vdots & \ddots & \vdots \\ -\gamma_{m1} & \cdots & \gamma_m - \gamma_{mm} \end{bmatrix}$$

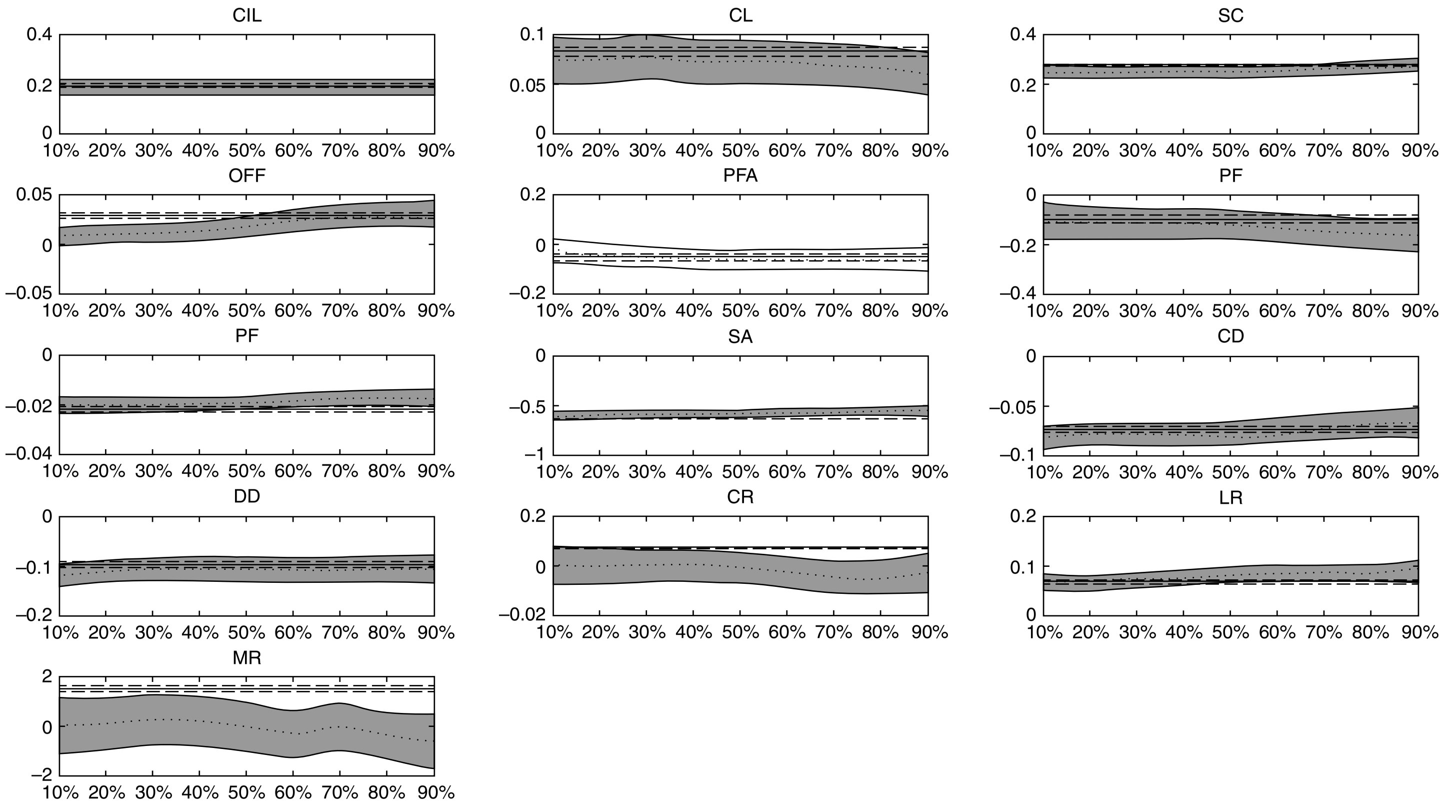


Figure 5.13 Panel data quantile regression elasticity estimates

$$+\begin{bmatrix} \delta_1\gamma_1+\theta_{11} & \cdots & \delta_n\gamma_1+\theta_{n1} \\ \vdots & \ddots & \vdots \\ \delta_1\gamma_m+\theta_{1m} & \cdots & \delta_n\gamma_m+\theta_{nm} \end{bmatrix} \begin{bmatrix} 2\delta_1^2+\delta_{11}-\delta_1 & \cdots & 2\delta_1\delta_n+\delta_{1n} \\ \vdots & \ddots & \vdots \\ 2\delta_n\delta_1+\delta_{n1} & \cdots & 2\delta_n^2+\delta_{nn}-\delta_n \end{bmatrix}^{-1}$$

$$\times\begin{bmatrix} \delta_1\gamma_1+\theta_{11} & \cdots & \delta_1\gamma_m+\theta_{1m} \\ \vdots & \ddots & \vdots \\ \delta_n\gamma_1+\theta_{n1} & \cdots & \delta_n\gamma_m+\theta_{nm} \end{bmatrix} \tag{5.15}$$

Based on sample measures, it is suggested that there is no evidence of economies of scope across all models among the three different types of loans evaluated at the sample mean point. Our results are consistent with the findings of Hughes and Mester (1993). They base their analysis on the translog cost dual in contrast to our primal output distance function. We both find no evidence of scale economies for the largest banks or significant scope economies. It is not clear that alternative nonparametric approaches such as the local linear approximations utilized by Wheelock and Wilson (2012) are directly comparable to our results, given their focus on banks of varying sizes and the substantial differences in number of parameters for such models. Constructing tests for the regularity conditions of the dual cost function from such innovative nonparametric approaches is a research issue that requires more study.

Figure 5.14 summarizes the results of the stochastic frontier estimation in terms of average efficiencies across the different estimators in each quarter. Efficiency levels range between about 0.10 and 0.4, using time-invariant estimators and with a downward trend using the BC model, whose specification requires that the temporal pattern is linear and monotonic: hence, the decline in average efficiency over the sample period from 75 percent to 70 percent. This trend is probably due to the substantial downturns in the recent period of the Great Recession and the financial meltdown.

The relationship between efficiency levels and bank sizes is also explored. From Figure 5.15 we can see that the largest banks do not necessarily have highest technical efficiencies; instead, the efficiency levels fluctuate as bank sizes change.

We further analyze the relationship between bank sizes and the Output Scale Efficiency ("OSE"). The derivation of this estimator follows Balk (2001).

$$OSE(x,y)=\frac{\overset{\vee}{D}{}_o^t(x,y)}{D_o^t(x,y)}=\frac{\overset{\vee}{OTE}{}_o^t(x,y)}{OTE_o^t(x,y)} \tag{5.16}$$

where the $\overset{\vee}{OTE}{}_o^t(x,y)$ is the output efficiency using cone technology (i.e., constant returns to scale – "CRS.") As we can see in Figure 5.16, which plots this OSE versus size ranking, the scale efficiencies estimated using time-invariant

Table 5.3 The scope economies estimates

	FE	RE	FEIV	REIV	H-T	PSS1	BC	QR(50%)
REL-CIL	0.0106	0.0108	0.0399	0.0278	0.0107	0.0212	0.0298	0.0160
	(0.0613)	(0.0267)	(0.0376)	(0.0430)	(0.0299)	(0.0751)	(0.0303)	(0.1134)
REL-CL	0.0266	0.0273	0.0307	0.0456	0.0267	0.0257	0.0237	0.0267
	(0.0400)	(0.0161)	(0.0292)	(0.0516)	(0.0144)	(0.0137)	(0.0147)	(0.0289)
REL-SC	0.0354	0.0322	0.0178	0.0037	0.0353	0.0360	0.0125	0.0290
	(0.0855)	(0.0157)	(0.0459)	(0.0781)	(0.0228)	(0.0212)	(0.0477)	(0.0450)
REL-OFF	0.0083	0.0125	0.0431	0.0601	0.0092	0.0131	0.0192	0.0120
	(0.0307)	(0.0050)	(0.0101)	(0.0581)	(0.0143)	(0.0359)	(0.0176)	(0.0094)
CIL-CL	–0.0030	–0.0034	–0.0368	–0.0502	–0.0030	–0.0094	0.0020	–0.0044
	(0.0907)	(0.0129)	(0.0321)	(0.0411)	(0.0143)	(0.0295)	(0.0219)	(0.0201)
CIL-SC	0.0514	0.0563	0.0677	0.0872	0.0522	0.0544	0.0625	0.0557
	(0.0401)	(0.0414)	(0.0457)	(0.0444)	(0.0335)	(0.0652)	(0.0219)	(0.0565)
CIL-OFF	0.0013	0.0020	–0.0391	–0.0533	0.0015	–0.0066	0.0043	–0.0098
	(0.0150)	(0.0097)	(0.0189)	(0.0266)	(0.0123)	(0.0210)	(0.0087)	(0.0226)
CL-SC	0.0273	0.0283	0.0073	0.0329	0.0277	0.0281	0.0002	0.0349
	(0.0268)	(0.0167)	(0.0395)	(0.0528)	(0.0214)	(0.0261)	(0.0293)	(0.0292)
CL-OFF	0.0055	0.0078	0.0029	–0.0011	0.0059	0.0066	0.0175	0.0053
	(0.0093)	(0.0069)	(0.0200)	(0.0171)	(0.0062)	(0.0110)	(0.0066)	(0.0129)
SC-OFF	–0.0054	–0.0085	0.0066	0.0161	–0.0061	–0.0063	–0.0157	–0.0017
	(0.0233)	(0.0088)	(0.0211)	(0.0157)	(0.0107)	(0.0171)	(0.0233)	(0.0039)

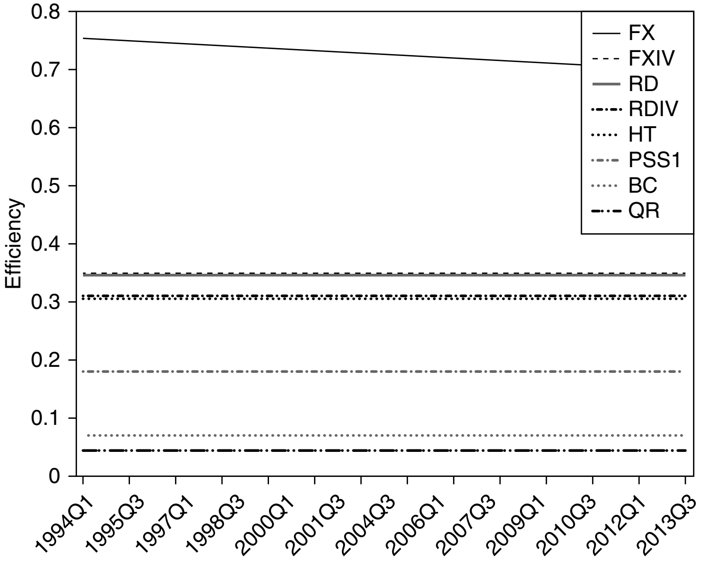


Figure 5.14 Estimated efficiencies using all stochastic frontier models

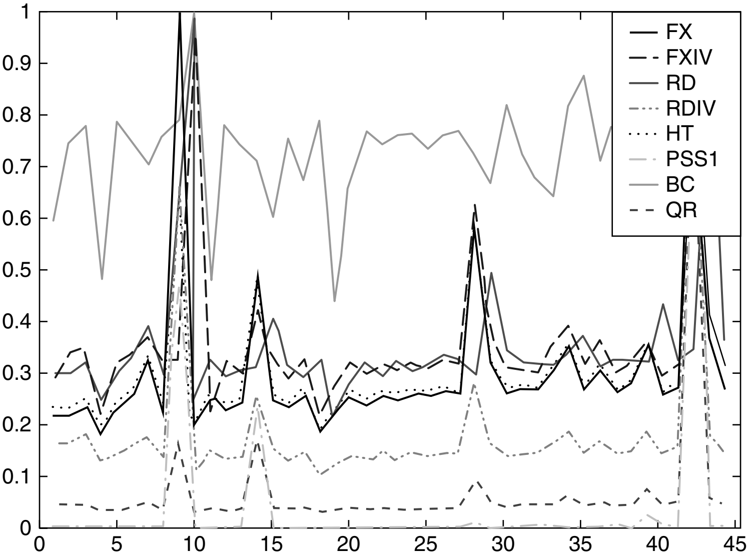


Figure 5.15 Efficiency levels and bank sizes

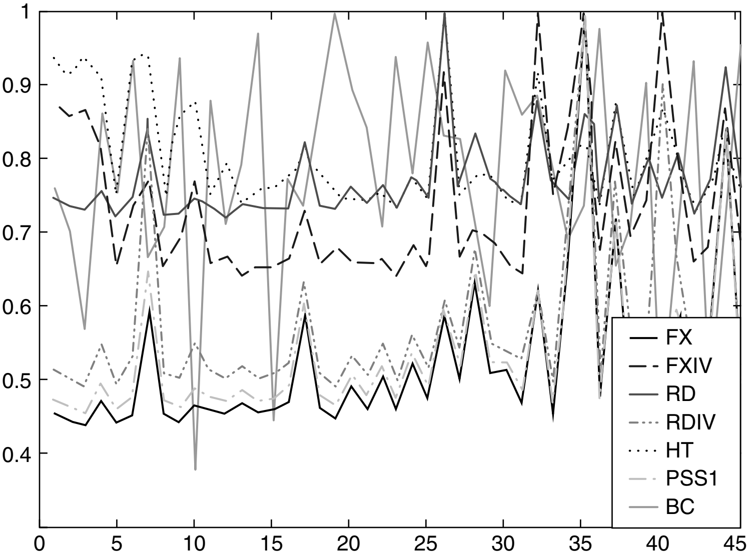


Figure 5.16 Scale efficiency plots using time-invariant estimators

estimators are increasing with fluctuations as bank sizes decrease (the ranking numbers increase). The scale efficiency level using the BC estimator,[15] although it displays a more fluctuating pattern than those using the time-invariant estimator, still suggests that large banks do not necessarily have higher scale efficiency levels.

5 Conclusion and directions for future research

This study represents a contribution to the recent dialogue that has arisen in the wake of the recent financial crisis, a re-examination amongst regulators, practitioners and academicians of the conventional wisdom regarding the implications of the spectacular growth of the financial sector of the economy. Previously, there was a widespread belief the "bigger is better," with arguments underpinning this view ranging from potential economies of scale and scope to a better competitive stance at the international level. We have seen this logic reversed in the post-crisis world to some degree, as for several reasons large banks have come to be viewed as a source of trouble and concern for policy makers and regulators.

We have addressed this controversy through an empirical analysis of the efficiency of U.S. banks with respect to their size and scope. This study utilized a new data set of bank history, a panel of financial measures derived from supervisory Call Reports in the period 1994–2013, from which we construct the variables used in both the frontier estimation and quantile regression analyses (inputs and outputs, as well as controls for three major risk types – credit, market and liquidity). In this exercise, we have been able to develop both policy implications and also evaluate potential analytical tools for supervisors.

The conclusion of the stochastic frontier estimation is that, in spite of growing, the largest U.S. surviving banks have decreased technical efficiency over the last 20 years. This has occurred as they took on increasing types of risk, and is reflected in an overall early decline in efficiency, as implied by the econometric model, which allows temporal variation. The estimation results also revealed no evidence on increasing returns to scale or scope across models. According to the time-invariant estimators, there is no positive correlation between bank size and technical efficiencies, neither does such a relationship exist between size and scale efficiencies. We found that credit, liquidity and market risks are deleterious to efficiency, which has implications for the argument that banks should be restricted to traditional banking activities in their zone of competence. The panel quantile regression results were generally consistent with the stochastic frontier estimation, albeit with estimates not varying greatly across quantiles. Furthermore, the implied efficiencies here are uniformly lower in the quantile regressions, than for the other time-invariant frontier estimators.

This paper has both policy implications and also evaluates various econometric techniques as potentially valuable analytical tools for supervisors. First, our results highlight the importance of the prudential supervisory role in controlling the level of risk in the banking sector (also reducing incentive for regulatory arbitrage between the banking and trading books), as we have documented that the elevation in risk measures, coupled with the growth of the sector, has resulted in declining measures of efficiency, a result that is robust to several econometric specifications. The policy implication is that we may want a better capitalized and somewhat smaller banking system, as this is likely to imply a more efficiently functioning banking industry. Second, the finding that market and liquidity risk dominate the influence of credit risk implied in the Volcker Rule debate, that regulators may wish not only to consider restricting banks from dangerous activities such as speculative proprietary trading, but also closely monitor the OTC exposures and their use for hedging some market risks instead of market-making purposes and consequently encourage insured commercial banks to focus on their core competency of making loans. There are several fruitful avenues of extension for this research program. We may pursue alternative data sets, such as other financial service types of firms (e.g., insurers, brokers), or data from other jurisdictions. We may expand our set of explanatory variables, with alternative controls (e.g., size, leverage, capitalization), or an expanded set of inputs (e.g., a measure of technological change.) Finally, we may expand our suite of alternative models, thereby seeking out further robust tools for use by supervisors.

Appendix

Table 5.4 Stochastic frontier estimates for translog distance function

Model	FE	RE	FEIV	REIV	H-T	PSS1	BC	Model	FE	RE	FEIV	REIV	H-T	PSS1	BC
CIL	0.210513	0.217183	0.225426	0.228280	0.211691	0.213949	0.226558	OFF*OFF	0.009649	0.010547	0.010252	0.011738	0.009810	0.005204	0.015968
	(0.007104)	(0.007128)	(0.008731)	(0.008980)	(0.007010)	(0.006812)	(0.007035)		(0.000893)	(0.000911)	(0.001082)	(0.001138)	(0.000883)	(0.001248)	(0.000941)
CL	0.081733	0.081414	0.049541	0.058110	0.081866	0.073736	0.070690	CIL*CL	0.003369	0.004041	0.016325	0.018603	0.003493	0.000263	0.005575
	(0.004743)	(0.004732)	(0.005757)	(0.005898)	(0.004673)	(0.004823)	(0.004959)		(0.001622)	(0.001655)	(0.001972)	(0.002072)	(0.001604)	(0.002309)	(0.001194)
SC	0.260411	0.270013	0.270424	0.282938	0.262164	0.267831	0.324163	CIL*SC	–0.008265	–0.008142	–0.034409	–0.033865	–0.008167	–0.016745	–0.006870
	(0.005294)	(0.005326)	(0.006756)	(0.006936)	(0.005224)	(0.005823)	(0.005665)		(0.002187)	(0.002206)	(0.002480)	(0.002587)	(0.002158)	(0.002876)	(0.002138)
OFF	0.044573	0.052108	0.029874	0.042833	0.046082	0.031973	0.075718	CIL*OFF	–0.005712	–0.006469	0.005667	0.005237	–0.005883	0.001587	–0.009611
	(0.003324)	(0.003347)	(0.004058)	(0.004210)	(0.003279)	(0.004090)	(0.003627)		(0.001939)	(0.001981)	(0.002227)	(0.002348)	(0.001918)	(0.002732)	(0.001910)
PFA	–0.048591	–0.051928	–0.090291	–0.087498	–0.050028	–0.043731	–0.025343	CL*SC	–0.015954	–0.018204	–0.010998	–0.017490	–0.016418	–0.013848	–0.024535
	(0.012755)	(0.012667)	(0.015723)	(0.016022)	(0.012551)	(0.011451)	(0.013190)		(0.001591)	(0.001587)	(0.001882)	(0.001927)	(0.001566)	(0.002146)	(0.001434)
NOE	–0.174543	–0.212133	–0.097825	–0.157074	–0.183900	–0.160083	–0.211568	CL*OFF	–0.005802	–0.006189	–0.011282	–0.012292	–0.005876	–0.003394	–0.008236
	(0.015366)	(0.014552)	(0.018810)	(0.018105)	(0.014950)	(0.014347)	(0.015879)		(0.001252)	(0.001277)	(0.001487)	(0.001565)	(0.001238)	(0.001779)	(0.001293)
PF	–0.021459	–0.020166	–0.022382	–0.020632	–0.021222	–0.022423	–0.014052	SC*OFF	0.014502	0.014889	0.022276	0.023005	0.014562	0.013018	0.016309
	(0.001488)	(0.001521)	(0.001797)	(0.001893)	(0.001473)	(0.002047)	(0.001649)		(0.001141)	(0.001167)	(0.001378)	(0.001452)	(0.001129)	(0.001641)	(0.001109)
SA	–0.551895	–0.558245	–0.545318	–0.552944	–0.553159	–0.561076	–0.552630	CIL*PFA	–0.051631	–0.051014	–0.071044	–0.078912	–0.051729	–0.059480	–0.013790
	(0.007726)	(0.007849)	(0.009602)	(0.010043)	(0.007637)	(0.009459)	(0.012772)		(0.006190)	(0.006295)	(0.005866)	(0.006128)	(0.006120)	(0.007656)	(0.006416)
CD	–0.058643	–0.056920	–0.044286	–0.042250	–0.058266	–0.060594	–0.071194	CIL*NOE	–0.013911	–0.010315	0.023763	0.039556	–0.013201	0.000593	–0.036446
	(0.003578)	(0.003631)	(0.004307)	(0.004495)	(0.003536)	(0.004021)	(0.004096)		(0.009364)	(0.009538)	(0.010167)	(0.010638)	(0.009260)	(0.011663)	(0.009503)
DD	–0.082796	–0.098362	–0.089356	–0.119685	–0.086103	–0.097145	–0.155330	CIL*PF	0.001301	0.000993	0.005045	0.005771	0.001252	–0.001339	–0.000405
	(0.006787)	(0.006797)	(0.008378)	(0.008589)	(0.006685)	(0.008305)	(0.007277)		(0.000744)	(0.000761)	(0.000804)	(0.000846)	(0.000736)	(0.001107)	(0.000816)
PFA*PFA	0.038638	0.033627	0.044351	0.025530	0.037510	0.041603	0.033637	CIL*SA	0.010498	0.009667	0.002173	0.003062	0.010466	0.012797	–0.008714
	(0.005790)	(0.005833)	(0.007338)	(0.007558)	(0.005712)	(0.007778)	(0.005474)		(0.003040)	(0.003098)	(0.004217)	(0.004416)	(0.003006)	(0.004296)	(0.002884)
NOE*NOE	–0.120394	–0.112475	–0.131007	–0.130757	–0.120410	–0.080091	0.142872	CIL*CD	0.019417	0.019489	0.015627	0.016128	0.019482	0.020563	0.015425
	(0.034828)	(0.035449)	(0.043828)	(0.045817)	(0.034436)	(0.045599)	(0.035533)		(0.001162)	(0.001182)	(0.001326)	(0.001391)	(0.001149)	(0.001664)	(0.001105)
PF*PF	–0.003953	–0.003604	–0.004079	–0.003435	–0.003887	–0.004157	–0.002120	CIL*DD	0.017048	0.017289	0.025161	0.024054	0.017052	0.016695	0.028778
	(0.000326)	(0.000333)	(0.000398)	(0.000418)	(0.000322)	(0.000448)	(0.000362)		(0.002334)	(0.002386)	(0.002795)	(0.002942)	(0.002309)	(0.003362)	(0.002465)
SA*SA	–0.064611	–0.062603	–0.064665	–0.062891	–0.064273	–0.065390	–0.035911	CL*PFA	0.006954	0.010037	0.027055	0.025806	0.007448	0.017517	0.034430
	(0.004188)	(0.004281)	(0.005258)	(0.005547)	(0.004143)	(0.006188)	(0.004678)		(0.004638)	(0.004733)	(0.005348)	(0.005593)	(0.004588)	(0.006404)	(0.004918)
CD*CD	–0.011441	–0.011021	–0.009875	–0.008825	–0.011356	–0.010620	–0.009391	CL*NOE	–0.022980	–0.028653	–0.046901	–0.049442	–0.023873	–0.050973	–0.074770
	(0.000803)	(0.000821)	(0.000953)	(0.001004)	(0.000795)	(0.001138)	(0.000897)		(0.008228)	(0.008385)	(0.009816)	(0.010261)	(0.008137)	(0.010956)	(0.008614)
DD*DD	–0.094605	–0.099422	–0.035313	–0.037091	–0.095586	–0.095532	–0.118533	CL*PF	–0.000451	–0.001304	–0.000003	–0.002056	–0.000643	–0.001116	–0.000478
	(0.005550)	(0.005674)	(0.005776)	(0.006087)	(0.005491)	(0.008130)	(0.004310)		(0.000563)	(0.000569)	(0.000582)	(0.000603)	(0.000556)	(0.000779)	(0.000597)
PFA*NOE	–0.135936	–0.128521	–0.140047	–0.099991	–0.134269	–0.142371	–0.154229	CL*SA	–0.031540	–0.028394	–0.058729	–0.057030	–0.030886	–0.015085	–0.020246
	(0.016387)	(0.016519)	(0.020987)	(0.021634)	(0.016168)	(0.021463)	(0.016586)		(0.002865)	(0.002920)	(0.003044)	(0.003203)	(0.002833)	(0.004066)	(0.002852)

continued

Table 5.4 Continued

Model	FE	RE	FEIV	REIV	H-T	PSS1	BC	Model	FE	RE	FEIV	REIV	H-T	PSS1	BC
PFA*PF	–0.003701	–0.004009	–0.005245	–0.006641	–0.003782	–0.002516	–0.002368	CL*CD	–0.000740	–0.000835	–0.000290	0.000709	–0.000768	0.000238	–0.001349
	(0.000781)	(0.000799)	(0.000938)	(0.000987)	(0.000773)	(0.001170)	(0.000849)		(0.000718)	(0.000735)	(0.000827)	(0.000872)	(0.000710)	(0.001070)	(0.000787)
PFA*SA	0.038080	0.027303	0.086320	0.058173	0.035773	0.035318	0.012125	CL*DD	0.003457	0.000906	0.030763	0.029668	0.002931	0.003483	–0.009432
	(0.009464)	(0.009575)	(0.011544)	(0.012023)	(0.009345)	(0.012147)	(0.008279)		(0.003038)	(0.003094)	(0.003183)	(0.003359)	(0.003003)	(0.004218)	(0.003030)
PFA*CD	0.002850	0.002147	–0.009586	–0.009082	0.002702	0.008438	0.017078	SC*PFA	0.000425	–0.007533	0.008197	–0.003508	–0.001080	0.008691	–0.051247
	(0.003410)	(0.003488)	(0.004087)	(0.004309)	(0.003374)	(0.004586)	(0.003790)		(0.004680)	(0.004747)	(0.005705)	(0.005973)	(0.004623)	(0.006436)	(0.004968)
PFA*DD	0.029591	0.034528	0.024848	0.039228	0.030696	0.031088	0.007854	SC*NOE	0.075593	0.089250	0.049030	0.060275	0.078081	0.045799	0.178468
	(0.009253)	(0.009406)	(0.011018)	(0.011565)	(0.009145)	(0.012356)	(0.009720)		(0.008372)	(0.008497)	(0.010034)	(0.010499)	(0.008273)	(0.011359)	(0.008942)
NOE*PF	0.012621	0.012775	0.012865	0.015652	0.012736	0.008392	0.005859	SC*PF	–0.000005	0.000169	0.000118	0.000366	0.000029	0.000658	0.000330
	(0.001966)	(0.002006)	(0.002362)	(0.002480)	(0.001944)	(0.002887)	(0.002101)		(0.000454)	(0.000465)	(0.000579)	(0.000611)	(0.000449)	(0.000686)	(0.000505)
NOE*SA	–0.010913	–0.003131	–0.076014	–0.053907	–0.008869	–0.003073	–0.057969	SC*SA	0.008717	0.003191	0.030313	0.023657	0.007587	0.012785	–0.014703
	(0.013275)	(0.013467)	(0.016498)	(0.017279)	(0.013115)	(0.017450)	(0.009738)		(0.002711)	(0.002712)	(0.003459)	(0.003629)	(0.002670)	(0.003547)	(0.002849)
NOE*CD	0.043213	0.042967	0.050378	0.043584	0.043351	0.038000	–0.004376	SC*CD	–0.017877	–0.017829	–0.019046	–0.018618	–0.017850	–0.016472	–0.018719
	(0.006301)	(0.006449)	(0.007477)	(0.007881)	(0.006235)	(0.008928)	(0.007085)		(0.001377)	(0.001411)	(0.001330)	(0.001403)	(0.001363)	(0.002068)	(0.001552)
NOE*DD	0.142434	0.139512	0.075760	0.063166	0.141948	0.119629	0.161181	SC*DD	–0.037741	–0.033453	–0.056338	–0.046868	–0.036754	–0.030332	–0.028529
	(0.012153)	(0.012380)	(0.014692)	(0.015396)	(0.012018)	(0.016112)	(0.011953)		(0.002806)	(0.002827)	(0.003632)	(0.003791)	(0.002768)	(0.003919)	(0.002981)
PF*SA	–0.000448	–0.001147	–0.000472	–0.002944	–0.000591	–0.000065	–0.001351	OFF*PFA	0.008623	0.008071	0.022027	0.022312	0.008503	0.008531	0.011195
	(0.001099)	(0.001125)	(0.001387)	(0.001458)	(0.001087)	(0.001640)	(0.001129)		(0.003513)	(0.003585)	(0.003905)	(0.004111)	(0.003475)	(0.004798)	(0.003778)
PF*CD	–0.000814	–0.000916	0.000260	–0.000244	–0.000842	–0.001108	–0.001098	OFF*NOE	–0.047185	–0.045010	–0.084294	–0.081532	–0.046567	–0.031327	–0.062669
	(0.000319)	(0.000326)	(0.000387)	(0.000407)	(0.000316)	(0.000478)	(0.000353)		(0.006275)	(0.006402)	(0.007415)	(0.007795)	(0.006206)	(0.008853)	(0.006622)
PF*DD	–0.001820	–0.001055	–0.001024	0.000747	–0.001703	0.002479	0.002019	OFF*PF	–0.002526	–0.002545	–0.003761	–0.004608	–0.002539	–0.001040	–0.002048
	(0.001284)	(0.001313)	(0.001488)	(0.001567)	(0.001271)	(0.001868)	(0.001336)		(0.000509)	(0.000522)	(0.000571)	(0.000601)	(0.000504)	(0.000773)	(0.000563)
SA*CD	0.014695	0.014271	0.022856	0.024622	0.014472	0.006685	0.031876	OFF*SA	0.044860	0.044592	0.052796	0.052436	0.044826	0.032180	0.047205
	(0.002513)	(0.002556)	(0.002994)	(0.003147)	(0.002484)	(0.003447)	(0.002667)		(0.003115)	(0.003186)	(0.003931)	(0.004149)	(0.003082)	(0.004556)	(0.003107)
SA*DD	0.011554	0.012161	0.007661	0.009862	0.011659	0.018797	0.013151	OFF*CD	–0.000669	–0.000900	0.006146	0.005435	–0.000782	–0.003741	0.007377
	(0.003989)	(0.004088)	(0.005196)	(0.005485)	(0.003948)	(0.006012)	(0.003455)		(0.001268)	(0.001288)	(0.001278)	(0.001342)	(0.001253)	(0.001781)	(0.001358)
CD*DD	–0.029968	–0.029502	–0.022906	–0.020797	–0.029892	–0.027892	–0.026060	OFF*DD	–0.000469	–0.001321	0.016510	0.016127	–0.000727	–0.001279	–0.002952
	(0.001945)	(0.001990)	(0.002304)	(0.002429)	(0.001925)	(0.002846)	(0.002052)		(0.002462)	(0.002514)	(0.003088)	(0.003250)	(0.002435)	(0.003594)	(0.002666)
CIL*CIL	0.013484	0.013819	0.010828	0.009638	0.013512	0.016347	0.017082	CR	–0.006039	–0.007250	–0.000928	–0.000464	–0.006253	–0.006200	–0.010448
	(0.002361)	(0.002408)	(0.002982)	(0.003129)	(0.002335)	(0.003451)	(0.002316)		(0.001962)	(0.002005)	(0.002382)	(0.002508)	(0.001941)	(0.002805)	(0.002161)
CL*CL	0.022843	0.026315	0.002024	0.012490	0.023612	0.023391	0.032332	LR	0.087475	0.087680	0.094706	0.092425	0.087282	0.101854	0.070947
	(0.001635)	(0.001583)	(0.001896)	(0.001872)	(0.001600)	(0.001957)	(0.001484)		(0.005511)	(0.005558)	(0.006857)	(0.007111)	(0.005440)	(0.006217)	(0.005836)
SC*SC	0.025020	0.023597	0.035500	0.033810	0.024727	0.032809	0.017752	MR	0.001398	0.000934	0.003520	0.002858	0.001295	0.000934	0.001112
	(0.001717)	(0.001735)	(0.002056)	(0.002156)	(0.001695)	(0.002430)	(0.001864)		(0.000840)	(0.000857)	(0.001027)	(0.001080)	(0.000831)	(0.001090)	(0.000923)

Table 5.5 Panel data quantile regression for translog distance function

Quantiles	10%	20%	30%	40%	50%	60%	70%	80%	90%	Quantiles	10%	20%	30%	40%	50%	60%	70%	80%	90%
CIL	0.186117	0.188241	0.186001	0.186772	0.181033	0.181395	0.180971	0.184430	0.188001	OFF*OFF	–0.000059	0.001693	0.002155	0.002412	0.003502	0.005009	0.005972	0.006503	0.006061
	(0.031082)	(0.030091)	(0.029643)	(0.029302)	(0.028715)	(0.028278)	(0.028355)	(0.028296)	(0.028207)		(0.004079)	(0.003808)	(0.003974)	(0.004252)	(0.004546)	(0.005026)	(0.005498)	(0.006088)	(0.006774)
CL	0.073422	0.072747	0.076757	0.072115	0.071939	0.071637	0.068294	0.066885	0.060639	CIL*CL	0.000443	–0.006342	–0.004294	–0.001453	0.002275	0.004204	0.007200	0.011679	0.017152
	(0.022942)	(0.022320)	(0.022035)	(0.021866)	(0.021725)	(0.021406)	(0.021272)	(0.021282)	(0.021927)		(0.019663)	(0.019261)	(0.019181)	(0.019501)	(0.019610)	(0.019913)	(0.019908)	(0.019667)	(0.018987)
SC	0.244247	0.242886	0.243926	0.247198	0.246216	0.248921	0.258118	0.261793	0.272753	CIL*SC	–0.006918	–0.012601	–0.020464	–0.022832	–0.022595	–0.025402	–0.031805	–0.033289	–0.040181
	(0.024862)	(0.023342)	(0.022519)	(0.022070)	(0.022037)	(0.022364)	(0.023317)	(0.024154)	(0.025485)		(0.019588)	(0.019370)	(0.019679)	(0.020014)	(0.020276)	(0.020441)	(0.020685)	(0.021068)	(0.021143)
OFF	0.007393	0.010903	0.011920	0.013627	0.018549	0.024395	0.028563	0.031227	0.031574	CIL*OFF	–0.004403	0.001008	0.008307	0.009789	0.007801	0.006278	0.005445	0.001246	–0.003303
	(0.009770)	(0.009068)	(0.009064)	(0.009533)	(0.010286)	(0.011022)	(0.011630)	(0.012374)	(0.013607)		(0.008668)	(0.008346)	(0.008292)	(0.008180)	(0.008112)	(0.008061)	(0.008062)	(0.008359)	(0.008583)
PFA	–0.028775	–0.041039	–0.053032	–0.062913	–0.063975	–0.062595	–0.062139	–0.065550	–0.062789	CL*SC	–0.012418	–0.005872	–0.010110	–0.013342	–0.016610	–0.018800	–0.027769	–0.033906	–0.043187
	(0.047097)	(0.042669)	(0.041194)	(0.040629)	(0.040481)	(0.040788)	(0.042346)	(0.043698)	(0.046825)		(0.016556)	(0.016966)	(0.017545)	(0.018072)	(0.018703)	(0.019232)	(0.019713)	(0.020280)	(0.021018)
NOE	–0.106631	–0.116144	–0.119710	–0.111422	–0.120446	–0.128904	–0.146938	–0.154808	–0.166384	CL*OFF	–0.000668	–0.000715	–0.001431	–0.001080	–0.001763	–0.002800	–0.003382	–0.003864	–0.006450
	(0.074282)	(0.067869)	(0.064095)	(0.061759)	(0.060331)	(0.060247)	(0.060722)	(0.062074)	(0.066665)		(0.008012)	(0.007594)	(0.007459)	(0.007479)	(0.007350)	(0.007119)	(0.006949)	(0.007060)	(0.007447)
PF	–0.020028	–0.020099	–0.019860	–0.019899	–0.019499	–0.018155	–0.017258	–0.017479	–0.017131	SC*OFF	0.009461	0.009532	0.005067	0.002478	0.004640	0.008444	0.010190	0.010825	0.013126
	(0.003180)	(0.002944)	(0.002840)	(0.002844)	(0.002902)	(0.003023)	(0.003108)	(0.003227)	(0.003644)		(0.009530)	(0.009285)	(0.009200)	(0.009330)	(0.009436)	(0.009538)	(0.009718)	(0.010129)	(0.010498)
SA	–0.611272	–0.610463	–0.596040	–0.595692	–0.590527	–0.586778	–0.576617	–0.571867	–0.560169	CIL*PFA	–0.043873	–0.059388	–0.059525	–0.056574	–0.043563	–0.051296	–0.050116	–0.050367	–0.055707
	(0.041616)	(0.039749)	(0.040038)	(0.040558)	(0.041477)	(0.041994)	(0.043100)	(0.045179)	(0.047135)		(0.043357)	(0.041435)	(0.041033)	(0.040350)	(0.039501)	(0.040048)	(0.040074)	(0.041054)	(0.042339)
CD	–0.084041	–0.078585	–0.080373	–0.080577	–0.077824	–0.076388	–0.071339	–0.069169	–0.067562	CIL*NOE	–0.012550	0.009259	0.009773	0.002164	–0.012836	0.003356	0.012006	0.028564	0.052372
	(0.011711)	(0.011349)	(0.011529)	(0.011649)	(0.012013)	(0.012269)	(0.012743)	(0.013527)	(0.014765)		(0.046693)	(0.043244)	(0.042900)	(0.042960)	(0.043771)	(0.045391)	(0.045088)	(0.045771)	(0.045582)
DD	–0.120876	–0.110563	–0.109319	–0.107017	–0.108603	–0.110255	–0.110080	–0.106796	–0.108408	CIL*PF	–0.003163	–0.001826	–0.001752	–0.001397	–0.001182	–0.000935	–0.001294	–0.000532	–0.000268
	(0.024460)	(0.022693)	(0.022591)	(0.023387)	(0.024175)	(0.024917)	(0.025462)	(0.026094)	(0.027727)		(0.002842)	(0.002495)	(0.002269)	(0.002143)	(0.002045)	(0.002003)	(0.001929)	(0.001952)	(0.002145)
PFA*PFA	0.004240	0.015394	0.009792	0.016539	0.029268	0.028861	0.039438	0.048291	0.055277	CIL*SA	0.028620	0.024623	0.025347	0.021883	0.021027	0.015455	0.006919	0.005040	0.002256
	(0.077997)	(0.071545)	(0.072075)	(0.073069)	(0.073929)	(0.076973)	(0.079013)	(0.085591)	(0.091551)		(0.026299)	(0.025230)	(0.024460)	(0.024340)	(0.024270)	(0.024293)	(0.024618)	(0.025038)	(0.025287)
NOE*NOE	–0.277984	–0.214868	–0.202672	–0.154166	–0.096597	–0.052156	–0.009614	0.073097	0.148470	CIL*CD	0.015550	0.015416	0.015467	0.015689	0.016193	0.015152	0.013273	0.009081	0.005077
	(0.220373)	(0.203961)	(0.196646)	(0.192185)	(0.189366)	(0.186824)	(0.181178)	(0.176364)	(0.182179)		(0.011954)	(0.011244)	(0.011282)	(0.011403)	(0.011391)	(0.011451)	(0.011253)	(0.011042)	(0.011149)
PF*PF	–0.003993	–0.003895	–0.003824	–0.003768	–0.003566	–0.003233	–0.002999	–0.003012	–0.002943	CIL*DD	0.004990	0.004088	0.006423	0.012573	0.013163	0.012411	0.015027	0.011620	0.007247
	(0.000689)	(0.000625)	(0.000606)	(0.000595)	(0.000594)	(0.000611)	(0.000637)	(0.000671)	(0.000718)		(0.025929)	(0.023831)	(0.023574)	(0.023213)	(0.023065)	(0.022882)	(0.022375)	(0.022070)	(0.023279)
SA*SA	–0.131778	–0.120157	–0.114132	–0.105562	–0.097685	–0.069441	–0.062722	–0.065735	–0.069211	CL*PFA	–0.019006	–0.013576	–0.007459	0.001497	0.003074	–0.002486	–0.004584	–0.003551	–0.000398
	(0.072298)	(0.067192)	(0.065457)	(0.066043)	(0.066695)	(0.066766)	(0.066100)	(0.066775)	(0.070128)		(0.031414)	(0.029947)	(0.029385)	(0.029059)	(0.029057)	(0.028808)	(0.028679)	(0.029095)	(0.032728)
CD*CD	–0.013915	–0.011969	–0.012407	–0.012991	–0.012969	–0.011669	–0.012190	–0.011839	–0.010899	CL*NOE	0.019910	0.006143	–0.019048	–0.039204	–0.048781	–0.044945	–0.037567	–0.038904	–0.039054
	(0.009021)	(0.008061)	(0.007827)	(0.007836)	(0.007761)	(0.007617)	(0.007437)	(0.007287)	(0.007473)		(0.042309)	(0.042427)	(0.043281)	(0.043549)	(0.043710)	(0.043103)	(0.042444)	(0.042487)	(0.044688)
DD*DD	–0.121485	–0.106573	–0.099866	–0.095124	–0.089777	–0.080274	–0.079112	–0.077630	–0.077686	CL*PF	–0.002419	–0.002389	–0.000128	0.000498	0.001072	0.001074	0.000764	0.000163	–0.000535
	(0.051855)	(0.050767)	(0.051501)	(0.052472)	(0.053469)	(0.054828)	(0.055517)	(0.054908)	(0.053177)		(0.002844)	(0.002468)	(0.002300)	(0.002172)	(0.002074)	(0.002035)	(0.002035)	(0.002009)	(0.002124)
PFA*NOE	–0.022658	–0.053140	–0.049211	–0.075778	–0.102124	–0.114556	–0.141830	–0.171989	–0.196975	CL*SA	–0.005841	–0.001207	0.007162	0.013905	0.015304	0.009243	–0.000491	–0.000983	–0.003212
	(0.108605)	(0.101312)	(0.097419)	(0.095836)	(0.095918)	(0.097020)	(0.098647)	(0.102838)	(0.112356)		(0.025980)	(0.026606)	(0.027167)	(0.028088)	(0.028291)	(0.028290)	(0.028721)	(0.029184)	(0.028938)

continued

Table 5.5 Continued

Quantiles	10%	20%	30%	40%	50%	60%	70%	80%	90%	Quantiles	10%	20%	30%	40%	50%	60%	70%	80%	90%
PFA*PF	–0.004135	–0.004058	–0.002634	–0.002188	–0.002461	–0.002494	–0.002940	–0.003035	–0.003332	CL*CD	–0.000894	0.000039	0.000943	0.000498	0.000125	0.001598	0.004394	0.004715	0.003923
	(0.005208)	(0.004887)	(0.004699)	(0.004509)	(0.004256)	(0.004184)	(0.004109)	(0.004139)	(0.004518)		(0.006783)	(0.006427)	(0.006374)	(0.006534)	(0.006557)	(0.006655)	(0.006560)	(0.006662)	(0.007095)
PFA*SA	0.002079	0.019333	0.025385	0.033430	0.044718	0.045205	0.059056	0.064307	0.063630	CL*DD	–0.018824	–0.016297	–0.010121	–0.007191	–0.002353	0.001986	0.000306	0.001579	0.000999
	(0.053450)	(0.049795)	(0.049898)	(0.050659)	(0.051505)	(0.052573)	(0.054353)	(0.055979)	(0.060246)		(0.019785)	(0.019021)	(0.018889)	(0.018671)	(0.019305)	(0.020057)	(0.021083)	(0.022048)	(0.022355)
PFA*CD	0.003788	0.004721	0.008706	0.009094	0.012963	0.018661	0.016394	0.015720	0.019363	SC*PFA	0.019211	0.020027	0.015906	0.000680	–0.005457	–0.000743	0.003554	0.008901	0.022205
	(0.019769)	(0.018817)	(0.018625)	(0.019118)	(0.019477)	(0.019891)	(0.020404)	(0.020557)	(0.022216)		(0.045221)	(0.044078)	(0.042227)	(0.040501)	(0.039057)	(0.039341)	(0.039082)	(0.039709)	(0.042169)
PFA*DD	0.008863	0.000462	–0.007166	–0.005463	–0.003192	0.008126	0.018582	0.040539	0.058235	SC*NOE	–0.003299	–0.009162	–0.003441	0.022859	0.038688	0.026509	0.010955	–0.001386	–0.029780
	(0.048706)	(0.046576)	(0.046981)	(0.048494)	(0.050594)	(0.052694)	(0.055732)	(0.059758)	(0.066622)		(0.056764)	(0.052546)	(0.050924)	(0.049832)	(0.049481)	(0.050550)	(0.051116)	(0.053818)	(0.058657)
NOE*PF	0.014510	0.013387	0.008623	0.006825	0.006088	0.006221	0.006425	0.009668	0.011091	SC*PF	0.001815	0.001803	0.001172	0.000161	0.000059	0.000404	0.000554	0.000737	0.000773
	(0.006977)	(0.006311)	(0.006091)	(0.006086)	(0.006078)	(0.006345)	(0.006469)	(0.006772)	(0.007302)		(0.001925)	(0.001869)	(0.001837)	(0.001824)	(0.001798)	(0.001741)	(0.001730)	(0.001754)	(0.001867)
NOE*SA	0.078369	0.045897	0.038290	0.024324	0.007162	–0.002470	–0.018505	–0.034500	–0.039404	SC*SA	0.006385	0.007168	0.009345	0.012282	0.008468	0.015464	0.031426	0.031101	0.035015
	(0.088053)	(0.077467)	(0.072146)	(0.068662)	(0.068247)	(0.069024)	(0.071582)	(0.071743)	(0.079383)		(0.027109)	(0.028519)	(0.029486)	(0.030086)	(0.030539)	(0.031121)	(0.031172)	(0.031623)	(0.033265)
NOE*CD	0.019795	0.017011	0.016829	0.022157	0.020355	0.007194	0.005859	–0.001592	–0.015464	SC*CD	–0.014223	–0.013682	–0.015866	–0.015984	–0.015563	–0.012504	–0.014253	–0.012262	–0.009196
	(0.031795)	(0.028458)	(0.027934)	(0.028245)	(0.028801)	(0.029313)	(0.029492)	(0.030700)	(0.033346)		(0.008641)	(0.007885)	(0.008067)	(0.008206)	(0.008540)	(0.008938)	(0.009246)	(0.009819)	(0.010974)
NOE*DD	0.123591	0.134359	0.126361	0.123569	0.111481	0.093693	0.084231	0.046737	0.016028	SC*DD	–0.012551	–0.003507	–0.003921	–0.013066	–0.020587	–0.023324	–0.018709	–0.020093	–0.021357
	(0.080366)	(0.081925)	(0.085897)	(0.089448)	(0.092691)	(0.095013)	(0.096408)	(0.097515)	(0.101100)		(0.028554)	(0.027142)	(0.025896)	(0.025575)	(0.026102)	(0.027207)	(0.027896)	(0.028225)	(0.027448)
PF*SA	0.000747	0.000178	0.001857	0.001914	0.001519	0.000632	–0.000940	–0.002889	–0.003419	OFF*PFA	0.003294	0.012839	0.007385	0.010693	0.009576	0.013579	0.013325	0.011411	0.008177
	(0.004379)	(0.004345)	(0.004178)	(0.004109)	(0.004035)	(0.004060)	(0.004022)	(0.004086)	(0.004291)		(0.018150)	(0.016673)	(0.016319)	(0.016519)	(0.016347)	(0.016503)	(0.016344)	(0.017043)	(0.019188)
PF*CD	–0.001207	–0.001049	–0.001170	–0.001205	–0.001178	–0.001389	–0.001005	–0.001175	–0.001038	OFF*NOE	0.002545	–0.016623	–0.011525	–0.014102	–0.018633	–0.028091	–0.032548	–0.034663	–0.032113
	(0.001013)	(0.000936)	(0.000934)	(0.000980)	(0.001005)	(0.001065)	(0.001087)	(0.001114)	(0.001232)		(0.027300)	(0.025821)	(0.026568)	(0.027319)	(0.027935)	(0.029788)	(0.030639)	(0.032301)	(0.033530)
PF*DD	–0.003948	–0.001787	–0.000101	0.001216	0.001898	0.001897	0.002506	0.002107	0.001603	OFF*PF	–0.000692	–0.001272	–0.000524	–0.000475	–0.000464	–0.000647	–0.000636	–0.001250	–0.001495
	(0.003914)	(0.003365)	(0.003068)	(0.002968)	(0.002864)	(0.002933)	(0.002962)	(0.003187)	(0.003382)		(0.001262)	(0.001141)	(0.001068)	(0.001031)	(0.001025)	(0.001047)	(0.001084)	(0.001123)	(0.001205)
SA*CD	0.009768	0.011469	0.006873	0.004952	0.003433	0.004085	0.008083	0.013691	0.016800	OFF*SA	0.016207	0.017883	0.013201	0.011523	0.016264	0.021786	0.024548	0.025219	0.024074
	(0.014931)	(0.014132)	(0.013970)	(0.014029)	(0.014227)	(0.014683)	(0.015016)	(0.015709)	(0.017652)		(0.013582)	(0.013421)	(0.014028)	(0.015146)	(0.015976)	(0.016910)	(0.018098)	(0.018827)	(0.019680)
SA*DD	0.042938	0.037763	0.039886	0.036899	0.036608	0.026969	0.026273	0.034153	0.040287	OFF*CD	–0.006953	–0.004950	–0.002379	–0.002060	–0.001670	–0.001899	–0.000278	0.001791	0.002455
	(0.038620)	(0.036002)	(0.035775)	(0.035673)	(0.036016)	(0.036479)	(0.036929)	(0.037907)	(0.039262)		(0.005829)	(0.005325)	(0.005301)	(0.005625)	(0.005956)	(0.006152)	(0.006302)	(0.006353)	(0.006809)
CD*DD	–0.019018	–0.019829	–0.017269	–0.018403	–0.017216	–0.011542	–0.010104	–0.006105	–0.002901	OFF*DD	–0.009685	–0.006710	–0.007389	–0.007009	–0.004731	–0.003444	–0.004073	–0.001233	0.003232
	(0.016251)	(0.014647)	(0.013986)	(0.013629)	(0.013911)	(0.014794)	(0.015903)	(0.016724)	(0.018251)		(0.014473)	(0.013035)	(0.013085)	(0.013052)	(0.013537)	(0.014686)	(0.015582)	(0.016524)	(0.016468)
CIL*CIL	0.023733	0.026440	0.024502	0.022840	0.022976	0.022139	0.019804	0.021462	0.032720	CR	0.000061	–0.000493	0.000201	0.000198	–0.001211	–0.002520	–0.004857	–0.005360	–0.003074
	(0.032839)	(0.031638)	(0.031057)	(0.031105)	(0.030785)	(0.030062)	(0.029174)	(0.028538)	(0.027727)		(0.007669)	(0.007143)	(0.006691)	(0.006466)	(0.006260)	(0.006209)	(0.006391)	(0.006773)	(0.007697)
CL*CL	0.020918	0.020436	0.033350	0.035086	0.037895	0.038761	0.041581	0.043548	0.048366	LR	0.064091	0.062788	0.072054	0.072946	0.082132	0.083271	0.086403	0.083422	0.090649
	(0.016586)	(0.017614)	(0.018079)	(0.019014)	(0.019929)	(0.020698)	(0.021746)	(0.023207)	(0.024633)		(0.016852)	(0.015036)	(0.014705)	(0.015034)	(0.015327)	(0.015670)	(0.015807)	(0.015993)	(0.016445)
SC*SC	0.034174	0.031780	0.040532	0.047459	0.044996	0.044263	0.060456	0.069094	0.085110	MR	–0.000035	–0.000015	0.000227	0.000147	–0.000036	–0.000402	–0.000064	–0.000455	–0.000639
	(0.036863)	(0.038031)	(0.039610)	(0.040700)	(0.041410)	(0.041901)	(0.042319)	(0.043457)	(0.044703)		(0.001135)	(0.001038)	(0.001011)	(0.000991)	(0.000981)	(0.000960)	(0.000940)	(0.000954)	(0.001097)

Table 5.6 Summary of econometric models

Model Name	Abbreviation	Description
Fixed Effects Model	FE	Traditional Fixed Effects model assuming that regressors are correlated with the effect term.
Random Effects Model	RE	Traditional Random Effects model assuming that regressors are uncorrelated with effect term.
Fixed Effects with Instrumental Variables	FEIV	Fixed Effects model with the right-hand-side endogenous variables replaced by the lagged variables as instruments.
Random Effects with Instrumental Variables	REIV	Random Effects model with the right-hand-side endogenous variables replaced by the lagged variables as instruments.
Hausman-Taylor Model	H-T	Hausman-Taylor model assuming that some of the regressors are correlated with the effect term while some are not.
Park-Sickles-Simar Model	PSS1	A semiparametric model assuming that a set of regressors is conditionally independent of the effect given the set of correlated regressors.
Battese-Coelli Model	BC	A stochastic frontier model assuming time-varying effects with the specification of exponential functional form.
Quantile Regression Model	QR	A quantile regression approach assuming that the coefficients of the regressors are dependent on the quantiles and the effect term is allowed to correlate with regressors.

Notes

1. Public Law 111–203 Dodd–Frank Wall Street Reform and Consumer Protection Act.
2. Feldman argued that "...I am skeptical that reforms focused on size per se will achieve their stated purpose of addressing TBTF; I have more confidence in reforms that identify and address features that produce spillovers in the first place ...."
3. They conclude that scale economies appear to increase with bank size for large banks from a standard model of bank production that does not control for any TBTF funding cost advantage, while using an adjustment for the price of debt using the implicit funding subsidy they find evidence of constant returns to scale and possible scale diseconomies for large banks.
4. Note that greater automation could imply greater operational risk, which is an implicit element of cost, but that is beyond the scope of the current empirical treatment.
5. Also see BCBS (2009b).
6. As of 3Q2013, the total assets of all U.S. Insured Commercial banks is $13.5 trillion.
7. Therefore, we estimate the distance function under both Cobb-Douglas and translog specifications. We will discuss only for the translog distance function, as those for the Cobb-Douglas are qualitatively comparable. These results are available on request.

8. Alternatives to the BC specification of time-varying heterogeneity, which has the same pattern but different intercepts for different firms, such as the Cornwell et al. (1990) estimator, required too much temporal variation in efficiency scores than the sample contained and we were unable to implement this estimator in our translog specification.
9. For a further discussion of this issue, showing the use of similar data in models for risk aggregation see Inanoglu and Jacobs (2009).
10. OTC derivatives are financial contracts, which derive their values from underlying assets and market conditions. OTC derivatives create counterparty credit risk due to the risk of insolvency of one party before the settlement of the transactions. It is very difficult – if not impossible – to incorporate counterparty credit risk measures in an efficiency framework as counterparty credit risk measures are forward looking and constructed from "exposure profiles." See Jacobs (2014) for regulatory requirements for counterparty credit risk measurement.
11. See Inanoglu, H., Jacobs, Jr., M., and Karagozoglu, A.K. (2014) for an impact analysis of Basel 2.5 on banks' regulatory capital for trading portfolios.
12. A second consultative document was published in October 2013. http://www.bis.org/publ/bcbs265.htm
13. We note the deficiency of VaR measure especially after the crisis, but VaR is still the industry standard in measuring market risk.
14. The linearity of covariate effects across different quantiles is consistent with the standard interpretation of technical efficiency in the stochastic frontier paradigm as a radial measure.
15. For the BC estimator, we use the average-over-time scale efficiency level.

References

Allen, F. and Santomero, A.M. 2001. "What Do Financial Intermediaries Do?" *Journal of Banking and Finance*, 25: 271–294.

Balk, B.M. 2001. "Scale Efficiency and Productivity Change." *Journal of Productivity Analysis,* 15: 159–183.

Balk, B.M. 2008. *Price and Quantity Index Numbers: Models for Measuring Aggregate Change and Difference*. Cambridge University Press, Cambridge.

Baltensperger, E. 1980. "Alternative Approaches to the Theory of the Banking Firm." *Journal of Monetary Economics,* 6: 1–37.

Battese, G.E., and Coelli, T.J. 1992. "Frontier Production Functions, Technical Efficiency and Panel Data: with Application to Paddy Farmers in India." *Journal of Productivity Analysis*, 3: 153–169.

BCBS. 2009a. Revisions to the Basel II market risk framework. Updated December 2010.

BCBS. 2009b. Principles for Sound Stress Testing Practices and Supervision. Consultative Paper, May (No. 155).

BCBS. 2012. Fundamental Review of the Trading Book, http://www.bis.org/publ/bcbs219.pdf

Bernanke, B. 2009. *Financial Reform to Address Systemic Risk*. Council on Foreign Relations, Washington, DC March 10, 2009.

Caves, D.W., Christensen, L.R. and Diewert, W.E. 1982. "The Economic Theory of Index Numbers and the Measurement of Input, Output, and Productivity." *Econometrica,* 50: 1393–1414.

Coelli, T. 2000. *On the Econometric Estimation of the Distance Function Representation of a Production Technology*. Center for Operations Research & Econometrics, Universite Catholique de Louvain.

Coelli, T., and Perelman, S. 1996. "Efficiency Measurement, Multiple-output Technologie and Distance Functions: With Application to European Railways." CREPP Working Paper 96/05, University of Liege.

DeYoung, R. 2010. *Scale Economies Are a Distraction*. The Region, Federal Reserve Bank of Minneapolis.

Efron, B., and Tibshirani, R. 1986. "Bootstrap Methods for Standard Errors, Confidence Intervals, and Other Measures of Statistical Accuracy." *Statistical Science,* 1: 54–75.

Feldman, R. 2010. *Size and Regulatory Reform in Finance: Important but Difficult Questions*. The Region, Federal Reserve Bank of Minneapolis.

Frye, J., and Pelz, E. 2008. BankCaR (Bank Capital-at-Risk): US Commercial Bank Chargeoffs. Working Paper, Federal Reserve Bank of Chicago (2008: 3).

Galvao, A.F. 2011. "Quantile Regression for Dynamic Panel Data with Fixed Effects." *Journal of Econometrics,* 164: 142–157.

Gregory, J. 2014. *Central Counterparties: Mandatory Clearing and Bilateral Margin Requirements for OTC Derivatives*. John Wiley & Sons.

Hajargasht, G., Coelli, T., and Rao, D. 2008. "A Dual Measure of Economies of Scope." *Economics Letters,* 100: 185–188.

Haldane 2009. "Banking on the State." http://www.bis.org/review/r091111e.pdf

Hausman, J.A., and Taylor, W.E. 1981. "Panel Data and Unobservable Individual Effects." *Econometrica*, 49: 1377–1398.

Hughes, J.P., and Mester, L.J. 1998. "Bank Capitalization and Cost: Evidence of Scale Economies in Risk Management and Signaling." *Review of Economics and Statistics,* 80: 314–325.

Hughes, J.P., and Mester, L.J. 2008. "Efficiency in Banking: Theory, Practice and Evidence." FRB of Philadelphia Working Paper No. 08–01.

Hughes, J.P., Mester, L.J., and Moon, C.G. 2001. "Are Scale Economies in Banking Elusive or Illusive?: Evidence Obtained by Incorporating Capital Structure and Risk-taking into Models of Bank Production." *Journal of Banking & Finance*, 25: 2169–2208.

Inanoglu, H., Jacobs, Jr. M., and Karagozoglu, A.K. 2014 (Spring). "Empirical Analysis of Bank Capital and New Regulatory Requirements for Risks in Trading Portfolios." *Journal of Fixed Income*, 23 (4): 71–88.

Jorion, P. 2002. "How Informative are Value-at-risk Disclosures?" *The Accounting Review,* 77 (4): 911–931.

Klein, L.R. 1953. *A Textbook of Econometrics*. Row, Peterson & Company, Evanston, IL.

Koenker, R. 1984. "A Note on L-estimates for Linear Models." *Statistics & Probability Letters,* 2: 323–325.

Koenker, R. 2004. "Quantile Regression for Longitudinal Data." *Journal of Multivariate Analysis*, 91: 74–89.

Koenker, R. 2009. "Quantreg: Quantile Regression." R package version 4.

Koenker, R., Bassett, G. Jr. 1978. "Regression Quantiles." *Econometrica*, 46: 33–50.

Mosteller, F. 1946. "On Some Useful 'Inefficient' Statistics." *The Annals of Mathematical Statistics,* 17: 377–408.

Newey, W.K. 1990. "Semiparametric Efficiency Bounds." *Journal of Applied Econometrics,* 5: 99–135.

Pagan, A., and Ullah, A. 1999. *Nonparametric Econometrics*. Cambridge University Press, Cambridge.

Park, B.U., Sickles, R.C., and Simar, L. 1998. "Stochastic Panel Frontiers: A Semiparametric Approach." *Journal of Econometrics,* 84: 273–301.

Park, B.U., Sickles, R.C., and Simar, L. 2003. "Semiparametric-efficient Estimation of AR (1) Panel Data Models." *Journal of Econometrics*, 117: 279–309.

Park, B.U., Sickles, R.C., and Simar, L. 2007. "Semiparametric Efficient Estimation of Dynamic Panel Data Models." *Journal of Econometrics*, 136: 281–301.

Park, B.U., and Simar, L. 1994. "Efficient Semiparametric Estimation in a Stochastic Frontier Model." *Journal of the American Statistical Association,* 89: 929–936.

Pitt, M.M., and Lee, L.F. 1981. "The Measurement and Sources of Technical Inefficiency in the Indonesian Weaving Industry." *Journal of Development Economics*, 9: 43–64.

R Core Development Team. 2010. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria.

Ritter, C., and Simar, L. 1997. "Pitfalls of Normal-gamma Stochastic Frontier Models." *Journal of Productivity Analysis,* 8: 167–182.

Schmidt, P., and Sickles, R.C. 1984. "Production Frontiers and Panel Data." *Journal of Business & Economic Statistics,* 2: 367–374.

Sealey, C.W., and Lindley, J.T. 1977. "Inputs, Outputs, and a Theory of Production and Cost at Depository Financial Institutions." *The Journal of Finance*, 32: 1251–1266.

Sickles, R.C. 2005. "Panel Estimators and the Identification of Firm-specific Efficiency Levels in Parametric, Semiparametric and Nonparametric Settings." *Journal of Econometrics*, 126: 305–334.

Tarullo, D.K. 2011. *Industrial Organization and Systemic Risk: An Agenda for Further Research.* Conference on Regulating Systemic Risk, Washington, DC September 15, 2011.

Tracey, B., and Davies, R. 2012. Too Big to be Efficient? The Impact of Implicit Funding Subsidies on Scale Economies in Banking. Working Paper.

Wheelock, D.C., and Wilson, P.W. 2012. "Do Large Banks have Lower Costs? New Estimates of Returns to Scale for US Banks." *Journal of Money, Credit and Banking*, 44: 171–199.

6
Efficiency, Competition and the Shadow Price of Capital

Thomas Weyman-Jones

1 Introduction and motivation

The purpose in writing this chapter encompasses several different motivations. The starting point is the study of recent developments in modeling banking systems from the point of view of the economics of industrial organization. This approach has gained attention in recent years due to the work of Freixas and Rochet (2008) and Degryse et al. (2009). There is a need to understand how efficiency can be evaluated in the banking system over the last decade. In addition it is useful to estimate the cost of equity capital, the primary loss-absorbing capacity of the banking system. This is difficult when some banks are not listed on the stock market, even though the necessity to raise equity capital remains important. Therefore a major motivation for the chapter is to suggest a model for estimating the shadow price of equity capital, or the shadow return on equity. These two aspects of efficiency and profitability come together in the measurement of competition in banking systems. This is a venerable topic in banking economics; but, in particular, the chapter examines recent developments in the literature, which bring together a measure of efficiency and a measure of profitability for the purpose of measuring the strength of competition.

Since a bank may experience a downgrade in the value of its assets, the role of equity is to provide the primary loss-absorbing capital that is required to maintain the solvency of the bank. There are costs and benefits to a financial stability policy that is aimed at increasing the capitalization of the banks in vulnerable times: while a greater equity capital buffer mitigates the riskiness of the balance sheet position, it is a costly policy, with the further drawback of being pro-cyclical in its macroeconomic impact. Usually the debate focuses on the idea that the market price of equity is higher than the price of debt, in this case deposits or borrowed funds. However, there have been several studies in the literature, notably Admati and Hellwig (2013) and Miles et al. (2011), which use Modigliani-Miller arguments amongst others to dispute the view that the case against re-capitalization can be made on the grounds of the higher weighted average cost of capital associated with equity financing. An alternative way of looking at the trade-off of costs and benefits of re-capitalization is to note that deleveraging to meet a policy constraint reduces the bank's profit-maximizing activity and therefore is measureable through the impact of the shadow price of the capital constraint on the bank's returns.

The chapter begins with an analysis of the problem of modeling the shadow return on equity capital in banking systems. It goes on to suggest how, in a formal model of cost-minimizing behavior, the idea of the dual cost function can be used to measure the decision-making that underpins a bank's optimizing behavior, subject to the familiar balance sheet constraint that loans and investments should be balanced by deposits, borrowed funds and equity capital. Using the results of a number of recent empirical studies, estimates are derived of the behavior of the shadow return on equity. The chapter then considers how efficiency and productivity analysis can be applied to banking systems to generate a measure of efficiency for both listed and non-listed banks, in order to measure the costs of the re-capitalization process. On the basis of this analysis of cost efficiency and the shadow return on equity capital, the chapter considers recent work on the measurement of competition in banking systems. This work is based on the idea that the relationship between profitability – as measured, for example, by the shadow return on equity – and cost efficiency changes systematically as the intensity of market competition and rivalry changes, so that the profitability–efficiency relationship can offer a direct measure of the strength of competition.

2 The shadow price of equity in deposit-taking financial institutions

We begin the analysis with an informal discussion of the essential ideas, as illustrated in Figure 6.1 which captures the balance sheet representation of banking industry technology.

Our basic model of deposit-taking financial institutions is represented by the balance sheet condition $L = B + E$, where the asset side is represented by loans, L, and the liabilities are deposits and borrowed funds, B, and the pure loss-absorbing capacity, by equity, E. In different states of the world different levels of equity are required to ensure expected balance sheet solvency, resulting in an expected banking technology represented by the probability weighted input requirement set $\sum_{j=1}^{J} \pi_j I^j \left(L^0\right)$, for which the efficient boundary can be written as the banking technology transformation function $F(B, E; L) = 0$, illustrated in Figure 6.1. This represents the expected solvency relationship between deposits

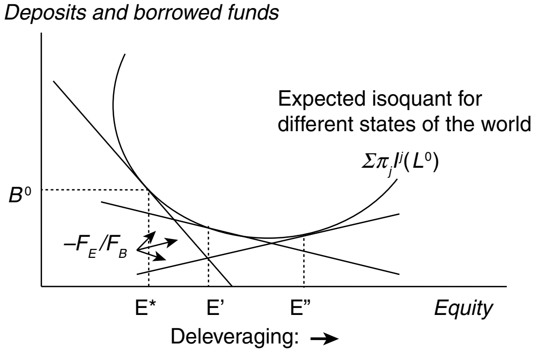

Deposits and borrowed funds
Expected isoquant for different states of the world
$\Sigma\pi_j l^j(L^0)$
B^0
$-F_E/F_B$
E*
E'
E"
Equity
Deleveraging: →

Figure 6.1 Representation of the balance sheet technology

and equity needed to support a given level of loans, L^0 in different states of the world $j = 1 \ldots J$ with probabilities: π_j. As attitudes to risk change, or there are changes in the risk associated with a given loan portfolio, a bank will aim for different equity levels to support the target level of loans, L^0 from a given level of deposits and borrowed funds: B^0. Three different equity levels are illustrated in the diagram, with the changes: $E* \rightarrow E' \rightarrow E''$representing the deleveraging required as there are increases in the perceived risk levels associated with a given balance sheet condition: $L^0 = B + E$. The technology is represented by the slope of the isoquant:

$$dB\big/dE = -F_E\left(B,E;L^0\right)\big/F_B\left(B,E;L^0\right) \tag{6.1}$$

When there is a minimal risk of default, the marginal rate of substitution is negative with positive shadow prices equated at the profit-maximizing position to the market interest rates on borrowed funds and loans:

$$-F_E\left(B,E;L^0\right)\big/F_B\left(B,E;L^0\right) = -r_E\big/r_B \tag{6.2}$$

As risk levels increase, the bank will require more loss-absorbing capacity to maintain the expected value of its balance sheet condition, and deleveraging will take place as the equity assets ratio is increased, either by raising equity capital or by calling in loans. The figure indicates that, as $E\big/L$increases, moving rightwards along the horizontal axis, the implied slope of the isoquant becomes less negative and may become positive if the technology displays weak disposability, representing the existence of an uneconomic region of the banking technology. Deleveraging leads to reductions in the shadow price of equity capital so that it may eventually become negative.[1] In addition, we note from the diagram that the allocative efficiency of equilibrium with other inputs may be disturbed by this phenomenon. It is argued in this chapter that this effect can be estimated by modeling the cost-minimizing behavior of the banking system, and we are able to do this by using the dual cost function to represent the technology. We capture the risk environment of the system by using risk based and market-based characteristics of the banks. Therefore we now develop a more formal model of cost-minimizing behavior.

3 Modeling the cost function and the shadow price of capital

In this section, we develop a cost-minimizing model of banking system activity that takes account of the equity capital requirements that must be met by banks and we show how increased capital requirements may impose additional costs on the efficient allocation of resources. The parametric frontier dual cost function that we will use is based on K variable inputs: $\mathbf{x} = (x_1, \ldots, x_K)$ with input prices: $\mathbf{w} = (w_1, \ldots, w_K)$ and R outputs: $\mathbf{y} = (y_1, \ldots, y_R)$, together with an additional quasi-fixed input, i.e., an input which may be a fixed input in the short run but is variable in the long run. For clarity, we symbolize this particular quasi-fixed input as z_0, with input price: w_0. The interpretation of this quasi-fixed input will be critical in the analysis of a banking industry sample since it captures the importance of the level of equity capital. We assume that this production technology has the properties of convexity, and weak disposability. It is the weak disposability assumption that is critical to our analysis. If the efficient boundary of the input requirement set at time t is represented by a transformation function: $F(\mathbf{y}, \mathbf{x}, z_0, t) = 0$ then weak disposability implies that the first derivatives, $F_k \equiv \partial F / \partial x_k$, $F_r \equiv \partial F / \partial y_r, F_z \equiv \partial F / \partial z_0$ are not restricted in sign. This will permit the model to accommodate both positive and negative shadow prices in the dual cost function. Adapting the arguments in Braeutigam and Daughety (1983) and Hughes et al. (2001), we write the dual long-run cost function, with all inputs including z_0 treated as variable, in the form:

$$c(\mathbf{y}, \mathbf{w}, w_0, t) = \min_{\mathbf{x}, z_0} \left\{ \mathbf{w}'\mathbf{x} + w_0 z_0 : F(\mathbf{y}, \mathbf{x}, z_0, t) = 0 \right\} \tag{6.3}$$

The short run cost function, on the other hand, with input z_0 treated as fixed, is the sum of variable and fixed cost:

$$c^s(\mathbf{y}, \mathbf{w}, \bar{z}_0, t) = c^v(\mathbf{y}, \mathbf{w}, \bar{z}_0, t) + w_0 \bar{z}_0 = \min_{\mathbf{x}} \left\{ \mathbf{w}'\mathbf{x} + w_0 z_0 : F(\mathbf{y}, \mathbf{x}, z_0, t) = 0; z_0 = \bar{z}_0 \right\} \tag{6.4}$$

The envelope theorem confirms that long-run total cost defines the envelope of short run total cost:

$$c(\mathbf{y}, \mathbf{w}, w_0, t) = \min_{z_0} \left\{ c^v(\mathbf{y}, \mathbf{w}, \bar{z}_0, t) + w_0 \bar{z}_0 \right\} \tag{6.5}$$

Consequently, the envelope theorem gives:

$$\partial c(\mathbf{y}, \mathbf{w}, w_0, t) / \partial z_0 = 0 = \left[\partial c^v(\mathbf{y}, \mathbf{w}, \bar{z}_0, t) / \partial z_0 \right] + w_0 \tag{6.6}$$

Rearranging this last result gives the critical interpretation of the shadow price of the quasi-fixed input:

$$-\left[\partial c^{v}\left(\mathbf{y},\mathbf{w},\bar{z}_0,t\right)\big/\partial z_0\right]=w_0 \tag{6.7}$$

This form of the envelope theorem is particularly relevant when, in addition to an input being fixed in the short run, there is no explicit information on its price. The negative of the derivative of the variable cost function with respect to this fixed input is the input's shadow price. This result becomes particularly useful when the input restriction arises because of regulatory intervention. Therefore, in modeling banking system cost we can think of z_0 as the regulated level of an input, equity capital, determined by the regulatory authority of the banking system. When expressed in logarithmic form the negative of the elasticity of cost with respect to equity capital: $-\left[\partial \ln c^{v}\left(\mathbf{y},\mathbf{w},\bar{z}_0,t\right)\big/\partial \ln z_0\right]= w_0 z_0\big/c=\varepsilon_{z0}$ is interpreted as the shadow return on equity. It is the share of total expenditure on inputs that accrues to equity owners when valued at the shadow price of equity capital. By including in a sample both listed and unlisted banks researchers are therefore able to calculate a shadow return on equity for the unlisted banks. This is therefore an important measure of profitability in banking.

Banks which are over-leveraged, or reliant on debt, and under-use equity capital can be expected to show a relatively high shadow return on equity (negative cost-elasticity with a relatively high absolute value), while banks which are less leveraged are likely to show a cost-elasticity with respect to equity that is lower in absolute value. Banks which are far from the long-run cost-minimizing equilibrium – for example, because they are undergoing major re-capitalization, with current equity capital levels well above the long-run equilibrium – may be expected to show a very low, possibly severely negative, shadow return on equity in the recovery phase from financial crisis. In general, negative values of the shadow input price or return on the fixed input would arise if, for example, the firm was operating in the uneconomic region of the production function.[2] The shadow return on the fixed input in the cost function measures the value of this possibly negative marginal product.

This approach to measuring profitability has been used in several applications to international banking systems. The original contribution of Hughes et al. (2001) used the shadow return on equity to adjust measures of economies of scale. Boucinha et al. (2013) used the shadow return on equity to analyze productivity developments and capital cost in the banking system in Portugal during the adoption of the Euro. They found that the shadow return on equity varied considerably as the effect of Portugal's membership of the Euro proceeded, reflecting changes in bank leverage. They used the approach as part of a wider study of productivity growth decomposition in the banking system in Portugal. Fethi et al. (2012) used the shadow return on equity to analyze productivity developments and capital cost in the banking system in Turkey, during the period of recovery from the financial crisis in that country in 2001. These authors discovered that the intense re-capitalization imposed on the banking system in Turkey, as part of the IMF restructuring and support program, led to the shadow return on equity turning negative for a period.

4 Measuring efficiency in banking systems

The actual variable cost experienced by the firm is by definition:$C_{it} \equiv \mathbf{w}'_{it}\mathbf{x}_{it}$, and consequently, the cost efficiency of bank i at time t is:

$$CE_{it} = \left\{ c^{s}\left(\mathbf{y}, \mathbf{w}, z_0, t\right)_{it} \big/ C_{it} \right\} \in (0,1] \tag{6.8}$$

Despite the immense amount of literature on measuring efficiency by stochastic frontier analysis and other methods, only a limited amount of attention has been paid to the question of whether an industry equilibrium of firms displaying different levels of efficiency is feasible. As we shall see below, it is feasible in a particular game theory framework which allows the measurement of relative efficiency – when it is combined with a measure of profitability, such as described in the previous section – to provide a very useful approach to measuring the strength of competition.

Using $\exp(-u), u \geq 0$ to transform the measure of cost efficiency from the interval: $(0,1]$ into a non-negative random variable with support on the non-negative real line: $[0,+\infty)$, yields:

$$\ln C_{it} = \ln c^{s}\left(\mathbf{y}, \mathbf{w}, z_0, t\right)_{it} + u_{it} \tag{6.9}$$

This can be modeled by a fully flexible functional form, such as the translog function, with an additive idiosyncratic error term v to capture sampling, measurement and specification error. Note that the cost function should be homogeneous of degree +1 in input prices which can be imposed by dividing through by one of the input prices, w_K. Define the variables in vector form as:

$$\begin{aligned} \mathbf{l}\tilde{w} &= \begin{pmatrix} \ln\left(w_1/w_K\right) & \ldots & \ln\left(w_{K-1}/w_K\right) \end{pmatrix} \\ \mathbf{ly} &= \begin{pmatrix} \ln y_1 & \ldots & \ln y_R \end{pmatrix} \end{aligned}$$

We write the translog approximation to (6.8) with additive error term as follows,[3]

$$\begin{aligned} \ln\left(C/w_K\right) = {} & \alpha_0 + \alpha'\mathbf{ly} + \beta'\mathbf{l}\tilde{\mathbf{w}} + \tfrac{1}{2}\mathbf{ly}'\mathrm{A}\,\mathbf{ly} + \tfrac{1}{2}\mathbf{l}\tilde{\mathbf{w}}'\mathrm{B}\,\mathbf{l}\tilde{\mathbf{w}} + \mathbf{ly}'\,\Gamma\,\mathbf{l}\tilde{\mathbf{w}} + \delta_1 t \\ & + \tfrac{1}{2}\delta_2 t^2 + \mu'\,\mathbf{ly}\,t + \eta'\,\mathbf{l}\tilde{\mathbf{w}}\,t + \rho_1 \ln z_0 + \tfrac{1}{2}\rho_2 \left(\ln z_0\right)^2 \\ & + \psi'\mathbf{ly}\ln z_0 + \xi'\mathbf{l}\tilde{\mathbf{w}}\ln z_0 + \phi \ln z_0 t + \mathbf{z}'\omega + v + u \end{aligned} \tag{6.10}$$

The vectors of elasticity functions (equivalent in the case of the input prices to the share equations by Shephard's lemma) are derived by differentiating the translog quadratic form:

$$\begin{bmatrix} \varepsilon_y \\ \varepsilon_{\tilde{w}} \\ \varepsilon_t \\ \varepsilon_{z0} \end{bmatrix} = \begin{bmatrix} \alpha & \mathrm{A} & \Gamma & \mu & \psi \\ \beta & \Gamma' & \mathrm{B} & \eta & \xi \\ \delta_1 & \mu' & \eta' & \delta_2 & \phi \\ \rho_1 & \psi' & \xi' & \varphi & \rho_2 \end{bmatrix} \begin{bmatrix} 1 \\ \mathbf{ly} \\ \mathbf{l}\tilde{w} \\ t \\ \ln z_0 \end{bmatrix} \tag{6.11}$$

The last line in this matrix equation recovers the critical elasticity of the cost function with respect to the quasi-fixed factor, i.e., the equity capital in the case of the banking system applications. It is the negative of this elasticity: -$\varepsilon_{z\mathrm{D}}$ that measures the shadow return on equity capital.

There are several options for estimation of this stochastic frontier analysis cost function depending on the specification adopted for the panel data composed error term: $v_{it} + u_{it}$ and most of these procedures are well known from the stochastic frontier analysis literature, e.g., Kumbhakar and Lovell (2003). Finally, we augment (6.9) by incorporating an additional vector of linear effects: $\mathbf{z}'\omega$ representing the risk characteristics of the environment in which the banks operate. These effects can be incorporated in two ways: directly in the parametric translog function in (6.9) so that the effects are instrumental in setting the location of the efficient frontier; or, alternatively, they can be used to control the parameters of the probability density function for the inefficiency component of the composed error term so that the effects determine the distance of any observation from the frontier.

There are many stochastic frontier analysis studies of efficiency in banking systems, but two that are particularly relevant to the topic of this chapter are the papers by Boucinha et al. (2013) and Fethi et al. (2012) already cited. Boucinha et al. (2013) used the model of stochastic frontier analysis of the cost function with the equity capital as a quasi-fixed input to develop a productivity decomposition of the banking system in Portugal during the first years of that country's membership of the Euro. Fethi et al. (2012) used a similar model to measure productivity developments in the banking system in Turkey in the recovery from the financial crisis of 2001. They discovered that, when the equity capital constraint is included in the productivity decomposition, the productivity growth in banking was reduced during periods of deleveraging because the shadow return on equity capital turned negative in those periods.

This chapter has described so far how the profitability and risks of the banking system are related to the return on equity capital, and how efficiency may be measured in banking systems. The next section brings together these two fundamental concepts to analyze the measurement of competition, a key aspect of banking regulation.

5 Measuring competition in banking systems

There are many models for measuring competition. Recent developments in this area include papers by Jan Boone et al. (2007) and Boone (2008). These are based on a Cournot model of competitive rivalry, and since both models have been applied to competition in banking systems, particularly in Europe, it is useful to develop them in detail.

The standard Cournot oligopoly game is well known. In the general case, with different marginal costs or different efficiencies for different entrants to the market, we have I firms competing as Cournot rivals where each has marginal cost: $c_i, i = 1 \ldots I$. Noting that the market price depends on the outputs of all firms together: $(q_i, \mathbf{q}_{-i})$ where (q_i) is the output of firm *i* and $(\mathbf{q}_{-i})$ is the vector of outputs of the other I$-$1 firms, the level of profit in equilibrium for firm *i* is:

$$\pi_i(c_i) = \max_{q>0}\{pq_i(c_i) - c_i q_i(c_i)\} = \max_{q>0}\{p(q_i, \mathbf{q}_{-i})\,q_i(c_i) - c_i q_i(c_i)\} \tag{6.12}$$

Then by a straight forward application of the envelope theorem: $d\pi_i(c_i)\big/dc_i = -q_i(c_i)$

i.e.,:

$$d\pi_i(c_i)\big/dc_i = \left[\left(q_i\left(dp(q_i, \mathbf{q}_{-i})\big/dq_i\right)\right) + p(q_i, \mathbf{q}_{-i}) - c_i\right] \times \left(dq_i\big/dc_i\right) - q_i(c_i) = -q_i(c_i) \tag{6.13}$$

The term in square brackets is the firm's marginal revenue – marginal cost condition taking the output of the other firms as given – and is equated to zero at the Cournot-Nash equilibrium.

Boone et al. (2007) and Boone (2008) develop this model in order to arrive at a test of competition. Boone et al. (2007) argue that more intense competition may have two forms: lower entry costs giving rise to a larger number of entrants and more aggressive rivalry amongst incumbents arising from anti-trust or regulatory intervention. In particular, they describe two possible effects of more competition: the *selection* effect whereby higher intensity of competition leads to inefficient firms exiting the market so that measured concentration appears to rise; and the *output reallocation effect* (call this ORE), whereby more competition causes output (or market share or profit) to shift relatively more to the most efficient or lowest cost firms in the market. If the most efficient firms have the highest price-cost margin, PCM, then the share-weighted industry average PCM will rise with increased competition. Consequently, the two traditional measures of competition – the Herfindhal-Hirschman index of market concentration, based on the sum of squared market shares, and the price-cost margin – may move in the counter-intuitive direction of increase when competition intensifies.

Boone et al. (2007) and Boone (2008) use the *output reallocation effect* (ORE: more competition causes output to shift relatively more to the most efficient firms in the market) to develop monotonic tests of the strength of competition based on the insight that "in a more competitive industry firms are punished more harshly for being inefficient" (Boone 2008, 1246). They have two different tests of competition; Boone et al. (2007) introduces the *profit elasticity*, PE: i.e., the percentage rise in the firm's profit for a one percent fall in its marginal cost, while Boone (2008) develops the *relative profit difference*, RPD: i.e., the relative profits of a more efficient firm compared to those of a less efficient firm when competition increases and the least efficient firm is the baseline comparator for both of the other firms.

When the firms are treated as producing differentiated products the demand curve of each firm reflects its differentiated product status and is written so that the relationship between the market price and that firm's own output has a different slope from the relationship between the market price and the output of its rivals, i.e., the goods are no longer perfect substitutes:

$$p\left(q_i, \mathbf{q}_{-i}\right) = a - bq_i - d\sum_{j\neq i}^{I-1} q_j, \text{ therefore for firm } i,: b = -\partial p\big/\partial q_i > 0;$$

$$d = -\partial p\big/\partial q_j > 0$$

The parameter *d* requires further analysis. In Boone et al. (2007, 8, 10) it is described as capturing the extent to which consumers see the different products in a market as close substitutes and he states that this is a common way of parameterizing competition in the literature.

If product differentiation gives firms some market power, then there is no direct competition between firms; making goods closer substitutes by raising the values of *d* towards *b* reduces this market power and intensifies competition. In Boone (2008, 1247) he refers to an analogous parameter as affecting the aggressiveness of each firm's conduct in the market, with higher values signifying more aggressive conduct by other firms, e.g., following a regulatory intervention such as a change in market structure brought about as a result of central bank oversight of a national banking system.

For firm *i*, the profit maximization problem is:

$$\max_{q>0} \pi_i = \left[a - bq - d\sum_{j\neq i}^{I-1} q_j \right] q - c_i q \tag{6.14}$$

With first order condition for each of the I firms:

$$d\pi_i/dq = a - 2bq_i - d\sum_{j\neq i}^{I-1} q_j - \left\{ qd \times \left(\frac{d\left(\sum\limits_{j\neq i}^{I-1} q_j\right)}{dq} \right) \right\} - c_i = 0 \quad i = 1\ldots I \tag{6.15}$$

The fourth term in the expression for the derivative is zero by the Cournot-Nash assumption that each firm takes the output of the other firms as given. In the special case where $d = b$, i.e., homogeneous products, the first order conditions are:

$$d\pi_i/dq = a - 2b\left(q_i + \sum_{j\neq i}^{I-1} q_j\right) - c_i = 0 \Rightarrow \left(q_i + \sum_{j\neq i}^{I-1} q_j\right) = \frac{a - c_i}{2b} \equiv s(c_i) \quad i = 1\ldots I \tag{6.16}$$

The industry Cournot-Nash equilibrium occurs at the simultaneous solution of all the reaction functions:$\mathbf{Rq} = \mathbf{s}(\mathbf{c})$ so that: $\mathbf{q} = \mathbf{R^{-1}s}(\mathbf{c})$. In the standard non-differentiated products case

$$\mathbf{R} = \begin{bmatrix} 1 & \frac{1}{2} & \cdots & \frac{1}{2} \\ \frac{1}{2} & 1 & \cdots & \frac{1}{2} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{1}{2} & \cdots & \frac{1}{2} & 1 \end{bmatrix} = \tfrac{1}{2}\left(\mathbf{I} + \mathbf{ii}'\right) \Rightarrow \mathbf{R}^{-1} = 2\left[\mathbf{I} - \left(\tfrac{1}{1+n}\right)\left(\mathbf{ii}'\right)\right] \tag{6.17}$$

Therefore, the non-differentiated product Cournot-Nash model supports an industry equilibrium in which different firms have different outputs depending on their different marginal costs, or efficiency levels.

For differentiated products, the corresponding result is:

$$\begin{bmatrix} 1 & \frac{d}{2b} & \cdots & \frac{d}{2b} \\ \frac{d}{2b} & 1 & \cdots & \frac{d}{2b} \\ \vdots & \ddots & \ddots & \vdots \\ \frac{d}{2b} & \cdots & \frac{d}{2b} & 1 \end{bmatrix} \begin{bmatrix} q_1 \\ q_2 \\ \vdots \\ q_I \end{bmatrix} = \begin{bmatrix} (a - c_1)/2b \\ (a - c_2)/2b \\ \vdots \\ (a - c_I)/2b \end{bmatrix} \text{ or}$$

$$\begin{bmatrix} 2b & d & \ldots & d \\ d & 2b & \ldots & d \\ \vdots & \ddots & \ddots & \vdots \\ d & \ldots & d & 2b \end{bmatrix} \begin{bmatrix} q_1 \\ q_2 \\ \vdots \\ q_I \end{bmatrix} = \begin{bmatrix} a - c_1 \\ a - c_2 \\ \vdots \\ a - c_I \end{bmatrix} \tag{6.18}$$

Therefore once again:

$\mathbf{Rq} = \mathbf{s} \Rightarrow \mathbf{q} = \mathbf{R}^{-1}\mathbf{s}$ but this time the matrix structure is:

$$\mathbf{R} = \begin{bmatrix} 1 & \frac{d}{2b} & \cdots & \frac{d}{2b} \\ \frac{d}{2b} & 1 & \cdots & \frac{d}{2b} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{d}{2b} & \cdots & \frac{d}{2b} & 1 \end{bmatrix} = \left(\left(1 - \left(\tfrac{d}{2b}\right)\right)\mathbf{I} + \left(\tfrac{d}{2b}\right)\mathbf{ii}'\right)$$

$$\Rightarrow \mathbf{R}^{-1} = \left(\frac{1}{1-\left(\frac{d}{2b}\right)}\right)\left[\mathbf{I} - \left(\frac{\left(\frac{d}{2b}\right)}{1+(I-1)\left(\frac{d}{2b}\right)}\right)(\mathbf{ii}')\right]$$

$$= \left(\frac{\frac{2b}{d}}{\left(\frac{2b}{d}-1\right)}\right)\left[\mathbf{I} - \left(\frac{1}{\left[\frac{2b}{d}+(I-1)\right]}\right)(\mathbf{ii}')\right] \qquad (6.19)$$

In the general case of I firms:

$$q_i = q(c_i) = \frac{\left(\frac{2b}{d}-1\right)a - \left(\frac{2b}{d}+I-1\right)c_i + \sum_{j=1}^{I} c_j}{(2b+d(I-1))\left(\frac{2b}{d}-1\right)} \quad i = 1 \ldots I \qquad (6.20)$$

The derivative of this function is:

$$\frac{dq_i}{dc_i} = q'(c_i) = \frac{2-\left(\frac{2b}{d}+I\right)}{(2b+d(I-1))\left(\frac{2b}{d}-1\right)} \quad i = 1 \ldots I \qquad (6.21)$$

Therefore, for a wide range of plausible parameter values including: $b > d$ we expect this derivative to be negative in sign. A rise in marginal cost will lower the firm's output. The firm's profits excluding fixed cost are therefore:

$$\pi_i = \pi(c_i) = \left[a - bq_i(c_i) - d\sum_{j \neq i}^{I-1} q_j(c_j)\right] q_i(c_i) - c_i q_i(c_1) \quad i = 1 \ldots I \qquad (6.22)$$

The comparative static effect is:

$$\frac{d\pi_i}{dc_i} = \pi'(c_i) = \left[\frac{\partial p}{\partial q_i} q_i(c_i) + p_i - c_i\right]\frac{dq_i}{dc_i} - q_i(c_i) = -q_i(c_i) < 0 \qquad (6.23)$$

As before, the term in square brackets is the firm's marginal revenue – marginal cost condition taking the output of the other firms as given, and is equated to zero at equilibrium, giving the envelope theorem result. Although this derivative is model and parameter dependent in sign, Boone argues that for a wide range of parameter values it will be negative and the slope will be steeper the higher the value of the parameter d, as illustrated in Figures 6.2 and 6.3. Both of these figures illustrate the *output reallocation effect*, ORE of competition in

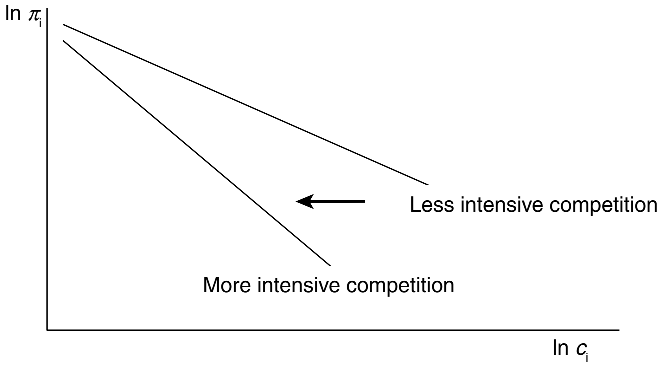


Figure 6.2 Boone et al. (2007) indicator of competitive pressure

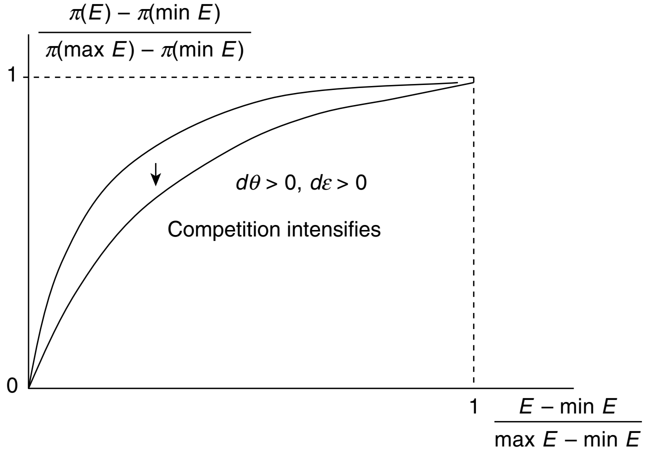


Figure 6.3 Boone (2008) RPD test of the intensity of competition

switching output from inefficient to efficient firms. The more intense is the competition, the greater is the reduction in the firm's profits for any given increase in its marginal cost, i.e., the more intense the competition the harsher the punishment for inefficiency. The Boone et al. (2007) indicator of competition intensity is the parametric size of this profit elasticity with respect to marginal cost.

Consequently, in Boone et al. (2007) the authors advocate the following test for competitive pressure: proceed by fitting, using dynamic panel data GMM procedures, a regression relating profitability to marginal cost and exogenous variables, with the elasticity of profitability with respect to marginal cost varying according to the period or competition regime in place, e.g.,

$$\ln \pi_{it} = \alpha - \beta_t \ln c_{it} + \mathbf{z'_{it}}\gamma + \varepsilon_{it} \tag{6.24}$$

The researcher may be able to compare estimates: $\hat{\beta}_t$ for $t = 1 \ldots T$ to determine different periods in which the degree of competitive pressure differed. Analogously, a similar relationship could be derived for both output and market share:

$$\ln q_{it} = \alpha - \beta_t \ln c_{it} + \mathbf{z'_{it}}\gamma + \varepsilon_{it} \tag{6.25}$$

$$\ln \left(q_i \Big/ \sum_j q_j \right)_{it} = \alpha - \beta_t \ln c_{it} + \mathbf{z'_{it}}\gamma + \varepsilon_{it} \tag{6.26}$$

In each case, the parameter $(-\beta_t)$ is expected to be negative and to increase in absolute value when the intensity of competition increases. Boone et al. (2007) applies this model in the form of (6.23) to the manufacturing sector in the Netherlands, while Van Leuvensteijn et al. (2011) apply it in the form of (6.26) to the loan markets of the Eurozone area banking systems and some other comparator countries for the period 1994–2004. Finally, Schaeck and Cihak (2014) estimate a similar model as part of their study of the relationship of competition to stability in banking systems. Van Leuvensteijn et al. (2011) find that there was considerable variation in the strength of competition amongst Eurozone banking systems in the period 1994–2004 prior to the financial crisis. They argue that this divergence was partly attributable to the characteristics of the different banking systems, since competition was stronger amongst commercial banks than savings banks due to the different degrees of openness to the global financial markets. The findings of Schaeck and Cihak (2014) produce broadly similar results for a panel dataset covering 1995–2005. Their estimates of the Boone profit elasticity vary across national EU banking systems in an approximate range from −0.25 to −1.5, and they showed some convergence as the period progressed.

Boone (2008) develops a similar idea, which he calls the *relative profit difference*, RPD. In this case, he distinguishes a measure of the firm's efficiency E_i so that higher efficiency shifts the total cost curve and the marginal cost curve down. Cost $C(q,E)$ has the properties: $C_q > 0, C_E \leq 0, C_{qE} \leq 0$, Boone (2008, 1247, assumption 1).

Consider the model used previously and replace the marginal cost term $c_i, i = 1 \ldots I$ with the measure of efficiency[4]: $E_i = \frac{1}{c_i}, i = 1 \ldots I$. Boone shows that a similar relation can be derived as before for the firm's optimum strategy, Boone

(2008, 1249, definition 1 and Equation (6.4)):

$$q_i = q(E_i) = \frac{\left(\frac{2b}{d} - 1\right) a - \left(\frac{2b}{d} + I - 1\right) \frac{1}{E_i} + \sum_{j=1}^{I} \frac{1}{E_j}}{\left(2b + d(I-1)\right)\left(\frac{2b}{d} - 1\right)} \qquad i = 1 \ldots I \tag{6.27}$$

Then:

$$\frac{dq_i}{dE_i} = q'(E_i) = \frac{\left(\frac{2b}{d} + I\right)\left(1/E_i^2\right)}{\left(2b + d(I-1)\right)\left(\frac{2b}{d} - 1\right)} > 0 \qquad i = 1 \ldots I \tag{6.28}$$

Write $\sum_{j=1}^{I} \frac{1}{E_j} \equiv E$ as a general index of efficiency, and generalize the intensity of competition parameter reflecting the price sensitivity to competing products to one reflecting the aggressiveness of firms' conduct in the market; this parameter, θ, increases the intensity of competition as it rises: it plays the role of d in the previous model: $d \equiv \theta$. An additional source of competition can be expressed by an increase in the parameter controlling a downward shift in entry costs: ε. Therefore more intense competition is represented as (i) lower entry costs: $d\varepsilon > 0$ or (ii) more aggressive inter-firm behavior: $d\theta > 0$.

Equation (6.28) above is related to the output reallocation effect, ORE, which Boone (2008) states as follows: after an increase in the intensity of competition, the increase in output of a more efficient firm exceeds the increase of a less efficient firm's output. There is a similar effect for $d\varepsilon > 0$. *Parameter changes:* $d\theta > 0$ (or $d\varepsilon > 0$) *will increase competition by raising RPD for any three firms with* $E^{**} > E^{*} > E$. That is,

$$\frac{d(RPD)}{d\theta} = \frac{d\left[\frac{\pi(E^{**}) - \pi(E)}{\pi(E^{*}) - \pi(E)}\right]}{d\theta} > 0 \tag{6.29}$$

The numerator is the cost reduction achieved by a firm with efficiency level E^{**} relative to a firm with the base efficiency level, E. The corresponding expression in the denominator is the cost reduction achieved by a firm with lower efficiency level E^{*} relative to a firm with the base efficiency level, E. Boone (2008) definition 2 states that stronger competition increases the cost advantage of the more efficient firm, and therefore that, as he puts it: "*in a more competitive industry firms are punished more harshly for being inefficient.*"

The key idea is contained in Boone (2008) theorem 1, and its explanation. Boone has established a relationship between the inverse relative profit difference,[5] which we will symbolize as ρ which he calls normalized profits and the corresponding normalized efficiency, symbolized here as η:

$$\rho = \left[\pi\left(E'\right) - \pi\left(\min E\right)\right] / \left[\pi\left(\max E\right) - \pi\left(\min E\right)\right] \tag{6.30}$$

$$\eta = \left[E' - \min E\right] / \left[\max E - \min E\right] \tag{6.31}$$

The relationship: $\rho(\eta)$must shift down for all values of the normalized efficiency when competition becomes more intense, Boone (2008: theorem 1). Boone suggests: plot normalized profits against normalized efficiency for the years t and t + 1. If the area under the curve is smaller in t + 1 than it is in t, competition has become more intense in year t + 1.

Using a diagram such as Figure 6.3 Boone represents an increase in competition intensity as a lower value for the integral under the curve: $\rho(\eta)$ i.e., $\int_0^1 \rho(\eta)\, d\eta$.

Boone's test is a sign criterion; in an analytical model the visual comparison of the areas under the relative profit difference graph, or the sign of their difference, is sufficient to determine the relative intensity of competition.

In deriving the test in this form, Boone has shown that in a sample of data, the sample points corresponding to the situation after competition has intensified will occupy a smaller space below the space occupied by the sample points for the situation before competition has intensified.

Duygun et al. (2015) used this approach to measure the shift in the RPD = ρ curve directly. They collect data from econometric analyses of banking systems costs in different emerging economies, including the transition economies that applied for entry to the European Union, in order to measure stochastic frontier analysis efficiency, E_{it}, and measured profitability, π_{it}. They used the shadow return on equity capital as the measure of profitability. They normalize the efficiency scores and the profitability measures in the form suggested by Boone. They then identify clusters of the sample points associated with different periods and different competition regimes. To implement the Boone test that more intense competition leads to a lower integral under the (inverse) RPD plot, Duygun et al. (2015) needed to estimate upper bounds for the clusters of the sample points corresponding to different periods or competition regimes. To avoid distortion by outliers while ensuring that a sufficiently large number of sample points is captured, they used polynomial quantile regressions to estimate the upper bounds and the areas under the curves for each cluster. Since the polynomial quantile regressions yield boundary estimates that depend on the quantile regression parameters, Duygun et al. (2015) are able to estimate the size of the Boone RPD integrals for different competition regimes and their standard errors, as well as to use Wald tests of the hypotheses that the intensity of competition has or has not changed from one period or regime to the next. Figure 6.4 illustrates the approach used in this test.

In Figure 6.4, an example is shown of two samples: A with points represented by shaded circles and B with points represented by unshaded circles. Each sample has outliers, which pull up the average regression line in sample A and pull down the average regression line in sample B. By fitting quantile regressions, for example at the third quartile, one can eliminate the impact of these outliers. The broken lines in the diagram represent these quantiles with 75 percent of

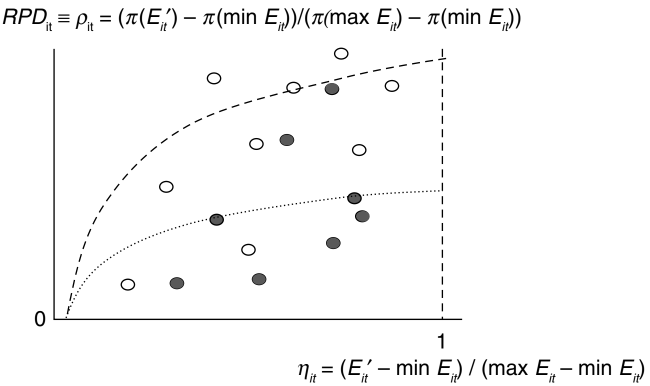


Figure 6.4 The sampled relationship between normalized profit (relative profit difference) and normalized efficiency with two quantile regression lines

Source: Duygun et al. (2015).

the respective sample points lying on or below the fitted quantile regressions. These quantile regressions also eliminate some of the apparent heteroscedasticity in the full sample. Since 75 percent of the sample points in sample cluster A occupy a smaller space below 75 percent of the sample points in sample cluster B, the conclusion is that competition has intensified.

The polynomial quantile regression model fitted at the third quartile is:

$$\Pr\left(\rho_{it} \leq \sum_{m=1}^{m=M} \alpha_m \eta_{it}^{m-1}\right) = q = 0.75 \tag{6.32}$$

In the quadratic case, with a dummy variable for sample A:

$$D_{it}^{A} = \begin{cases} 0, i, t \in B \\ 1, i, t \in A \end{cases} \tag{6.33}$$

Then the quadratic quantile regression line that splits the sample into clusters is

$$\rho_{it} = \alpha_1 + \alpha_2 \eta_{it} + \alpha_3 \eta_{it}^2 + \beta_1 D_{it}^A + \beta_2 \left(\eta_{it} \times D_{it}^A\right) + \beta_3 \left(\eta_{it}^2 \times D_{it}^A\right) + \varepsilon_{it} \tag{6.34}$$

A Wald test can be used assuming that

$$\varepsilon_{it} \sim N\left(0, \sigma^2\right) \tag{6.35}$$

Applied to the coefficients of the dummy variable, this is used to evaluate the statistical significance of the intensification of competition.

When these Boone tests are applied to panel data on banking systems, striking results emerge. In Duygun et al. (2015) the relative profit difference indicator is used to analyze the preparation of banking systems in the new members of the Eurozone under the guidelines of the European Central Bank in preparation for exposure to wider capital markets. This is shown to be associated with increased strength of competition. This preparation for Euro-convergence and convergence of banking systems occurred in the period 2000–2008. Duygun et al. (2015) find that using the Boone RPD test, the changes in the integral areas shown in Figure 6.4 are consistent with increased competition amongst these countries' banking systems, in preparation for the exposure to more globalized banking.

6 Conclusions and directions for future research

The purpose of this chapter can be stated quite simply. It is, firstly, to analyze two empirically measurable building blocks in the analysis of banking systems, i.e., the shadow return on equity capital as a measure of profitability and the stochastic frontier analysis of the banking system cost functions as a measure of efficiency; and, secondly, to combine these two empirical quantities, profitability and efficiency, into measures of the relative strength of competition in banking systems. I began the chapter by outlining a model of the equity capital and borrowed funds technology for a bank meeting a target loan portfolio in different states of the world. I formalized these ideas in a dual short run cost function for a weakly disposable technology that permits the existence of a non-economic region of the production set, treating the level of equity capital as the fixed or regulated input. I demonstrated that the negative of the elasticity of cost with respect to the fixed input equity level can be interpreted as the shadow return on the equity input by applying the envelope theorem. Several applications of this idea to banking systems were summarized. I then showed how to estimate this dual cost function model using a stochastic frontier analysis setting with maximum likelihood estimation applied to a panel dataset.

I then showed how the two building blocks represented by measured profitability and measured efficiency can be used together to measure the strength of competition in banking systems in different periods and different places. Following the analyses of Boone, I described both the profit elasticity test and the relative profit difference test of the relative strength of competition. Each shows that, because of the output reallocation effect, whereby more efficient banks are better able to maintain profitability than less efficient banks when there is an exogenous competition-enhancing shock to the relationship between measured profitability and measured efficiency, profit elasticity and relative profit difference can be interpreted as measures of changes in the strength of competition.

Finally, I reported on studies of European and emerging economy banking systems that applied these profit elasticity and relative profit difference tests to measuring the strength of competition in banking systems. The profit elasticity test studies showed that the strength of competition varied across banking systems depending on the relative importance of the commercial and savings banking sectors, while the relative profit difference study showed that competition in banking systems in potential Eurozone members had been enhanced by preparation for exposure to the global capital market.

Consider now some possible directions for future research. In this chapter, it has been argued that three key empirical aspects of the industrial organization approach to the analysis of banking systems are efficiency, profitability (as measured by the shadow return on equity capital) and competition, and it has been shown that these are closely related because it is possible to use the relationship between measured profitability and measured efficiency to measure the relative strength of competition.

Each of these areas offers considerable scope for analytical and empirical development. In the case of efficiency measurement, a fundamental concern is the researcher's ability to distinguish inefficiency from heterogeneity in stochastic frontier analysis. In truth, the core of the debate is that statistically there is no such thing as "inefficiency" or "heterogeneity," there are only random variables with probability distributions. Therefore, "inefficiency" or "heterogeneity" are interpretative concepts imposed on particular decompositions of regression residuals. If one can decompose the residual into two random variables, one of which is symmetrically distributed and which can be treated as idiosyncratic error, and one of which is asymmetrically distributed, then economic reasoning can be used to interpret the asymmetrically distributed random variable as "inefficiency" – provided the asymmetry is in the right direction, i.e., positively skewed for cost data. Greene (2005) however argues thus: suppose one can decompose the panel data residual into three random variables – one of which is symmetrically distributed and time varying and which can be treated as idiosyncratic error, one of which is symmetrically distributed and time invariant so that economic reasoning can be used to say it can be treated as heterogeneity, and one of which is asymmetrically distributed and time varying – then economic reasoning can be used to interpret the asymmetrically distributed time varying random variable as "inefficiency." In fact, part of the stochastic frontier analysis literature that distinguishes inefficiency and idiosyncratic error is simply a recasting of the original one-way panel data econometrics which distinguished heterogeneity and idiosyncratic error. Therefore, the development of robust models of cost functions with this three-way decomposition of the residuals is a key area for ongoing research. An alternative way in which to handle heterogeneity, idiosyncratic error and inefficiency has been suggested by the stochastic non-parametric envelopment of data approach of Kuosmanen et al. (2015). In this form of stochastic frontier analysis, the residual is subjected to the usual two-way decomposition into inefficiency and idiosyncratic error using parametric probability density functions, while the potential heterogeneity amongst economic agents is addressed by using a non-parametric formulation of the kernel behavioral function in the form of convex non-linear least squares, which permits each agent to have its own form of the cost function without assuming that cost-elasticity functions have the same parametric form across all sample observations, as is currently done in standard stochastic frontier analysis.

Turning to the return on equity capital and the role of equity in banking regulation, the comment by the former Governor of the Bank of England is apposite. Prior to the financial crisis, as King describes, rising returns on equity capital are often observed; but these are likely to have been due to increased leveraging by banks and reliance on short term funding. King goes on to note that, in the aftermath of the 2007–2008 financial crisis, the monetary authorities in Switzerland required their major banks to aim for equity–capital ratios far in excess of the Basel requirements and approaching 19–20 percent, just as the banking system in Turkey was doing after their financial crisis of 2001–2002. King's broad answer to the banking crisis is very simple: "much, much more equity; much, much less short term debt," King (2010:18). However, major re-capitalization of the banking systems around the world must impose resource costs, both on the wider economy and on the banking system in particular. Therefore, it is critical to include the role of equity capital as a quasi-fixed input in the short run cost function model of banks, or to include the price of equity capital in a long-run cost model since without either of these variables the cost functions will suffer from omitted variable bias.

The third issue that was covered was measurement of competition, a perennial topic in banking system analysis but one made even more relevant by the liberalization of the banking sector in both developed and emerging economies. For example, the Independent Banking Commission in the UK recommended that both divestment and increased competition through new entry should be policy priorities in the period following the recovery from the financial crisis of 2008. I argued that the innovative work of Jan Boone is critical in this respect. While there are many models for measuring competition, Boone showed that some of the most widely used could be ambiguous. His use of the generalized output reallocation effect allowed Boone to develop unambiguous and robust measures of the relative strength of competition. It is important to recognize that the Boone approaches – profit elasticity and relative profit difference – do not try to measure whether a particular market at a given date falls into one of the economic boxes labeled: monopoly, monopolistic competition, competition and so on. Instead, starting from a well-founded model of game theoretic behavior, Boone derives reduced-form relationships, which shift over time and across groups of banks as the competition regime becomes relatively stronger or weaker. The shift in these reduced-form relationships can be econometrically tested and robust conclusions drawn about whether a particular market at a particular time has experienced a shock equivalent to deregulation or new entry. This is likely to be a fruitful approach but the initial testing procedures described in this chapter are still relatively unproven, need further study and many more case study examples.

Finally, the chapter noted the Schaeck and Cihak (2014) paper, which brings together the topic of competition studied here with research on the financial stability of banking systems. It is interesting in this regard that the Bank of England has established a Financial Policy Committee to operate alongside its Monetary Policy Committee in a two-pronged attack on problems of banking system oversight and regulation. The link between competition and financial stability is critical and research on it is still at an early stage. One area of concern is that the conventional measures of stability concentrate on the relative volatility of market rates of return. The shadow rate of return on equity could also play an important role here, and dynamic measures of stability have hardly been explored at all.

Notes

This chapter is partly based on a research carried out with two colleagues, Professor Franco Fiordelisi (Durham) and Dr Nemanja Radic (Middlesex) and it draws on some ideas presented in a paper with the same title presented at *Financial Modelling Post-2008: Where Next? Seminar 2 – Distributional Assumptions and Efficiency*, School of Management, University of St Andrews on March 21, 2013.

1. This does not imply that the market cost of equity will ever become negative since the shadow price of equity is a lower bound for the market price.
2. The translog specification described in this chapter was specifically developed in order to allow operation in the uneconomic region of the technology; see Kumbhakar and Lovell (2003, 45).
3. Panel subscripts *it* are suppressed here for convenience.
4. In line with the standard notation in stochastic frontier analysis the symbol for efficiency used here is *E*, replacing Boone's symbol *n*.
5. The inverse relative profit difference is used to permit a diagrammatic interpretation and avoid division by zero.

References

Admati, Anat, and Martin Hellwig. 2013. *The Banker's New Clothes: What's Wrong with Banking and What to Do about It*. Princeton University Press, Princeton and Oxford.

Boone, Jan. 2008. "A New Way to Measure Competition." *The Economic Journal,* 118, (August): 1245–1261.

Boone, Jan, Jan C. van Ours, and Henry van der Wiel. 2007. "How (not) to Measure Competition." *Centre for Economic Policy Research Discussion Paper Series no. 6275*. London, and www.cepr.org

Boucinha, Miguel, Nuno Ribeiro, and Thomas Weyman-Jones. 2013. “An Assessment of Portuguese Banks’ Efficiency and Productivity Towards Euro Area Participation.” *Journal of Productivity Analysis,* 39 (2): 177–190.

Braeutigam, R.R., and Daughety, A.F. 1983. “On the Estimation of Returns to Scale Using Variable Cost Functions.” *Economics Letters,* 11: 25–31.

Degryse, Hans, Moshe Kim, and Steven Ongena. 2009 *Microeconometrics of Banking: Methods, Applications and Results*. Oxford University Press, Oxford.

Duygun, Meryem, Mohamed Shaban, and Thomas Weyman-Jones. 2015. “Measuring Competition using the Boone Relative Profit Difference Indicator.” *Economics Letters,* 132: 117–120.

Fethi, Meryem Duygun, Mohamed Shaban, and Thomas Weyman-Jones. 2012. “Turkish Banking Recapitalization and the Financial Crisis: An Efficiency and Productivity Analysis.” *Emerging Markets Finance and Trade,* 48 (sup5): 76–90.

Freixas, Xavier, and Jean-Charles Rochet, 2008. *Microeconomics of Banking* (2e). MIT Press, Cambridge, Mass.

Greene, William. 2005. “Reconsidering Heterogeneity in Panel Data Estimators of the Stochastic Frontier Model.” *Journal of Econometrics*, 126: 269–303.

Hughes, J.P., Mester, L.J., and Moon, C-G. 2001. “Are Scale Economies in Banking Elusive or Illusive?: Evidence Obtained by Incorporating Capital Structure and Risk Taking into Models of Bank Production.” *Journal of Banking and Finance*, 25.

King, M. 2010. “Banking from Bagehot to Basel and Back Again,” The Second Bagehot Lecture, Buttonwood Gathering, New York, October 25, 2010.

Kumbhakar, S., and Lovell, C.A.K. 2003. *Stochastic Frontier Analysis*. Cambridge, Cambridge University Press.

Kuosmanen, T., Johnson, A.L. and Saastamoinen, A. 2015. “Stochastic Nonparametric Approach to Efficiency Analysis: A Unified Framework,” In J. Zhu (ed.) *Data Envelopment Analysis: A Handbook of Models and Methods*. New York, Springer.

Miles, David, Jing Yang, and Gilberto Marcheggiano. 2011. “Optimal Bank Capital.” *Discussion Paper 31: Revised and Expanded Version* (April 2011), External MPC Unit, Bank of England.

Schaeck, Klaus, and Martin Cihák. 2014. “Competition, Efficiency, and Stability in Banking.” *Financial Management,* 43 (1): 215–241.

Van Leuvensteijn, Michael van, Jacob A. Bikker, Adrian, A.R.J.M., van Rixtel and Christopher Kok Sørensen. 2011. “A New Approach to Measuring Competition in the Loan Markets of the Euro Area.” *Applied Economics,* 43 (23): 3155–3167.

7
Model-Free Methods in Valuation and Hedging of Derivative Securities

Mark H. A. Davis

1 Introduction

Were "the quants" to blame for the financial crisis of 2008? In narrow terms the answer appears to be "no" on the argument put forward by Alex Lipton, that the banks that survived were using the same models as those that failed. Be that as it may, it does seem that a contributory factor in the crisis was over-reliance on models that, in retrospect, had insufficient credibility.

Modeling financial data is not simple: there is a great deal of it, and the stylized features of financial time series – heavy tails, irregular sample paths and stochastic volatility – make analysis difficult. In particular, direct prediction of prices, index values etc. is essentially impossible. The author's short paper (Davis 2016) on classifying prediction problems puts finance at the extreme end of the scale. In contrast to weather forecasting, another area where prediction is hard, we have no physical model for economic data, so prediction cannot be done using models with a clear scientific basis, and uncertainty can be reduced only by diversification of portfolios. The fact that the fund management industry employs quite so many people deploying such a bewildering variety of techniques is testament to the intractability of the problem.

There is another area of finance, however, where the degree of success has been much greater, namely the pricing and hedging of derivative securities. This topic was initiated by Louis Bachelier in his PhD thesis (Bachelier 1900) but, from the economic point of view, lay dormant for 65 years before being taken up by Paul Samuelson in the 1960s (see Samuelson 1965). Samuelson's approach, following Bachelier, of modeling prices as exogenously-defined stochastic processes was regarded as anathema by many economists of the day, who reasonably thought that the business of economists was explaining why prices are what they are, not just doing statistics. They missed the fact that the objective, more modest but more achievable, was not to explain the price of IBM stock but to explain the relationship between this price and the prices of options written on the stock. For this, the stochastic approach is the right way, as was definitively shown by Fischer Black and Myron Scholes in their great paper on option pricing (Black and Scholes 1973). This paper led to an explosion of activity, both in mathematics and in trading, over the next 25 years, leading to a complete theory of arbitrage pricing (Delbaen and Schachermayer 2008), a whole repertoire of price models and computational techniques (Glasserman 2003, Hull 2011, Hilber et al. 2013) and a massive expansion of investment banking making use of this new technology. The historical developments are traced in Davis and Etheridge (2006).

The classic approach in option pricing consists of the following three steps when the option in question has exercise value $\Phi(S)$ at time T where $S=\{S(t), 0 \leq t \leq T\}$ is the underlying price path:

(i) Select a class of price models $\mathfrak{M}$ whose sample paths and statistical properties are appropriate to the problem at hand. A model $M(\theta)$ within this class will be completely specified by a finite-dimensional parameter vector θ.
(ii) Using market interest rate and traded option data, calibrate the model, i.e., select a parameter vector $\hat{\theta}$ such that the model $M(\hat{\theta})$ minimizes the mean-squared error between model and market prices of the traded options.
(iii) Calculate the option value $p_\Phi = \mathbb{E}^{\hat{\theta}}[e^{-rT}\Phi(S)]$ as the risk-neutral discounted expected payoff of the option under the model $M(\hat{\theta})$.

The model will also determine hedge parameters but these are less directly useful as options are generally hedged on a book rather than individual basis. Essentially, this process is a kind of interpolation procedure: we produce a model $\hat{\theta}$ and a price p_Φ that is consistent with the market in that trading at that price alongside the existing traded options does not introduce arbitrage within that model. How successful this process is depends on the problem. Suppose for example that our traded options are call options maturing at times $T_1 < T_2 < \cdots < T_N = T$. If $\Phi(S) = \phi(S_{T_i})$ for some continuous function ϕ then because of the Breeden-Litzenberger formula (Section 3.1 below) all well-calibrated models will give essentially the same value for Φ. Now consider the case where Φ is a path-dependent option with exercise value $\Phi(S) = [X_T - K]^+$ where X_T is a weighted average, $X_T = \sum_1^N a_i S_{T_i}$. Then the traded options are missing essential information: their value only depends on the marginal distributions at the corresponding exercise time whereas to price a path-dependent option we need to know something about the joint distribution of prices at different times. We may therefore expect the path-dependent option price to be significantly model-dependent. The situation is if anything worse when Φ is a one-touch option that pays \$1 if $\max_{t \leq T} S_t \geq b$, where $b > S_0$, and zero otherwise. The traded options convey very little information about sample function behavior and we will get completely different answers depending on, for example, whether or not we allow for possible jumps in the sample functions of our model.

In recent times, and particularly since the 2008 crisis, it has been seen as essential to quantify the degree of uncertainty in option pricing occasioned by the variety of models used. A natural question to ask is: *What can we say about the value of an option with payoff Φ given* only *current market data, i.e., prices of underlying assets, yield curve data and prices of traded options?* This is the subject of the present chapter.

We start in Section 2 by reviewing the Black-Scholes formula and the underlying theory that makes it work. We draw attention in Section 2.3 to the still under-appreciated fact that the Black-Scholes model has a certain "robustness" in that the delta-hedging strategy can provide successful hedging even when the true price process is substantially different from the log-normal diffusion assumed by the model.

The starting point in extracting information about the true price process from option data is the *volatility surface*, a plot (with or without interpolation) of the Black-Scholes implied volatility as a function of the exercise time T and strike K of traded call or put options. This is the subject of Section 3. Very early on it was noted by Breeden and Litzenberger (1978) that the slice of the volatility surface at fixed T determines the risk-neutral probability distribution of S_T, the underlying asset price at time T. Later, Dupire (1994) showed that, with interpolation, the volatility surface actually determines a complete "local volatility model" whose marginal distributions must of course coincide with those given by Breeden and Litzenberger. These early contributions, summarized in Sections 3.1 and 3.2, form the basis for much of the subsequent work in this area.

Recent contributions divide into two categories, depending on whether we interpolate and extrapolate the given options data to give a complete volatility surface defined for all times and strikes $(t,k) \in [0,T] \times \mathbb{R}^+$ (where T is the last exercise time for which data is available), or whether we deal directly with the finite set $\hat{\sigma}_{ij}$ of implied volatilities at exercise times T_i and strikes K_{ij}. In the former case, discussed in Section 4, we can appeal to Breeden and Litzenberger (1978) and assume that marginal distributions are known at all exercise times T_i, whereas in the latter case (Section 5) we have to consider, in principle, all distributions consistent with the given data. The general objectives are (i) to determine the range of values for the price of some non-traded option that are consistent with absence of arbitrage, and (ii) to determine "semi-static" trading strategies that enable such options to be hedged.

Some concluding comments will be found in Section 6.

2 The Black-Scholes formula

It is no exaggeration to say that the history of financial economics divides sharply into two periods, the pre- and post-Black-Scholes eras, divided by the appearance of the paper Black and Scholes (1973) that initiated the modern theory of option pricing. The idea is revealed in the opening sentences of the abstract:

> If options are correctly priced in the market, it should not be possible to make sure profits by creating portfolios of long and short positions in options and their underlying stocks. Using this principle, a theoretical valuation formula for options is derived.

This encapsulates the basic idea, which is that – with the asset price model they employ – insisting on absence of arbitrage is enough to obtain a unique value for a call option on that asset. The resulting formula, (7.3) below, is certainly the most famous formula in financial economics.

2.1 The model and formula

Let $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t\in\mathbb{R}^+}, \mathbb{P})$ be a probability space with a given filtration $(\mathcal{F}_t)$ representing the flow of information in the market. Traded asset prices are $\mathcal{F}_t$-adapted stochastic processes on $(\Omega, \mathcal{F}, \mathbb{P})$. We assume that the market is *frictionless*: assets may be held in arbitrary amount, positive and negative, the interest rate for borrowing and lending is the same, and there is no bid–ask spread. While there may be many traded assets in the market, we fix attention on two of them. Firstly, there is a "risky" asset whose price process $(S_t, t \in \mathbb{R}^+)$ is assumed to satisfy the stochastic differential equation

$$dS_t = \mu S_t dt + \sigma S_t dW_t \tag{7.1}$$

with given *drift* μ and *volatility* σ. Here $(W_t, t \in \mathbb{R}^+)$ is an $(\mathcal{F}_t)$-Brownian motion. Equation (7.1) has the unique solution

$$S_t = S_0 \exp\left((\mu - \frac{1}{2}\sigma^2)t + \sigma W_t \right). \tag{7.2}$$

Asset S_t is assumed to have a constant *dividend yield* q, i.e., the holder receives a dividend payment $qS_t dt$ in the time interval $[t, t+dt]$. Secondly, there is a riskless asset paying interest at a fixed continuously-compounding rate r, whose value at $t \leq T$ is

$$B_t = \exp(-r(T-t)).$$

A *European call option* on S_t is a contract, entered at time 0 and specified by two parameters (K, T), which gives the holder the right, but not the obligation, to purchase one unit of the risky asset at price K at time $T > 0$. If $S_T \leq K$ the option is worthless and will not be exercised. If $S_T > K$ the holder can exercise his option, buying the asset at price K, and then immediately selling it at the prevailing market price S_T, realizing a profit of $S_T - K$. Thus the exercise value of the option is $[S_T - K]^+ = \max(S_T - K, 0)$. Similarly, the exercise value of a *European put option*, conferring on the holder the right to *sell* at a fixed price K, is $[K - S_T]^+$. In either case the exercise value is non-negative and, in the above model, is strictly positive with positive probability, so the option buyer should pay the writer a premium to acquire it. Black and Scholes (1973) showed that there is a unique arbitrage-free value for this premium.

Theorem 1. *(a) In the above model, the unique arbitrage-free value at time $t < T$ when $S_t = S$ of the call option maturing at time T with strike K is*

$$C(t,S) = e^{-q(T-t)}SN(d_1) - e^{-r(T-t)}KN(d_2) \tag{7.3}$$

where $N(\cdot)$ denotes the cumulative standard normal distribution function

$$N(x) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x} e^{-\frac{1}{2}y^2}\,dy \tag{7.4}$$

and

$$d_1 = \frac{\log(S/K) + (r - q + \sigma^2/2)(T-t)}{\sigma\sqrt{T-t}}, \tag{7.5}$$

$$d_2 = d_1 - \sigma\sqrt{T-t}.$$

(b) The function $C(t,S)$ may be characterized as the unique $C^{1,2}$ solution of the Black-Scholes partial differential equation (PDE)

$$\frac{\partial C}{\partial t} + (r-q)S\frac{\partial C}{\partial S} + \frac{1}{2}\sigma^2 S^2\frac{\partial^2 C}{\partial S^2} - rC = 0 \tag{7.6}$$

solved backwards in time with the terminal boundary condition

$$C(T,S) = [S-K]^+. \tag{7.7}$$

(c) The value of the put option with exercise time T and strike K is

$$P(t,S) = e^{-r(T-t)}KN(-d_2) - e^{-q(T-t)}SN(-d_1). \tag{7.8}$$

The theorem is proved by showing that the call option value can be replicated by a dynamic trading strategy investing in the asset S_t and in the riskless asset. A trading strategy is specified by an initial capital x and a pair of processes Δ_t, β_t representing the number of units of S, B respectively held at time t; the portfolio value at time t is then $X_t = \Delta_t S_t + \beta_t B_t$, and by definition $x = \Delta_0 S_0 + \beta_0 B_0$. The portfolio value is given by

$$X_T = x_0 + \int_0^T \Delta_u\,dS_u + \int_0^T \beta_u\,dB_u + \int_0^T q\Delta_u S_u\,du,$$

where the first integral is an Itô stochastic integral. The trading strategy is *self-financing* if

$$\Delta_t S_t + \beta_t B_t - \Delta_s S_s - \beta_s B_s = \int_s^t \Delta_u \, dS_u + \int_s^t q\Delta_u S_u \, du + \int_s^t \beta_u \, dB_u,$$

implying that the change in value over any interval in portfolio value is entirely due to gains from trade (the accumulated increments in the value of the assets in the portfolio plus the total dividend received).

It turns out (see Hull (2011) or Davis (2010) for the complete story) that replication is achieved, i.e., $X_T = [S_T - K]^+$ when

$$x = C(0, S_0), \quad \Delta_t = \frac{\partial C}{\partial S}(t, S_t), \quad \beta_t = \frac{1}{rB_t}\left(\frac{\partial C}{\partial t} + \frac{1}{2}\sigma^2 S_t^2 \frac{\partial^2 C}{\partial S^2} - qS_t \frac{\partial C}{\partial S}\right)$$

where C is given by (7.3). The main hedge parameter is the *Black-Scholes delta*, the number of units of the underlying asset in the hedge portfolio X_t, given explicitly by

$$\Delta_t = e^{-q(T-t)} N(d_1).$$

Key insights into the Black-Scholes formula are obtained by introducing a change of measure on the underlying probability space $(\Omega, \mathcal{F}_T, \mathbb{P})$. Define a measure $\mathbb{Q}$, the so-called *risk-neutral measure* by the Radon-Nikodým derivative

$$\frac{d\mathbb{Q}}{d\mathbb{P}} = \exp(-\theta W_T - \frac{1}{2}\theta^2 T).$$

Expectation with respect to $\mathbb{Q}$ will be denoted $\mathbb{E}_{\mathbb{Q}}$. By the Girsanov theorem, $\check{W} = W_t + \theta t$ is a $\mathbb{Q}$-Brownian motion, so that from (7.1) the SDE satisfied by S_t under $\mathbb{Q}$ is

$$dS_t = (r - q) S_t \, dt + \sigma S_t \, d\check{W}_t \tag{7.9}$$

so that for $t < T$

$$S_T = S_t \exp\left((r - q - \frac{1}{2}\sigma^2)(T - t) + \sigma(\check{W}_T - \check{W}_t)\right). \tag{7.10}$$

Applying the Itô formula we find that, with $\tilde{X}_t = e^{-rt} X_t$ and $\tilde{S}_t = e^{-rt} S_t$,

$$d\tilde{X}_t = \Delta_t \tilde{S}_t \sigma \, d\check{W}_t. \tag{7.11}$$

Thus (under technical conditions) $e^{-rt} X_t$ is a $\mathbb{Q}$-martingale. Let $h(S) = [S - K]^+$ and suppose there exists a replicating strategy, i.e., a strategy (x, Δ, β) with value process X_t such that $X_T = h(S_T)$ a.s. Then $\tilde{X}_t$ is a $\mathbb{Q}$-martingale, and hence for $t < T$

$$X_t = e^{-r(T-t)} \mathbb{E}_{\mathbb{Q}}[h(S_T) | \mathcal{F}_t] \tag{7.12}$$

and in particular

$$x = e^{-rT} \mathbb{E}_{\mathbb{Q}}[h(S_T)]. \tag{7.13}$$

We see that x only depends on the one-dimensional distribution of S_T. From (7.10), this is a log-normal distribution. Writing $(\check{W}_T - \check{W}_t) = Z\sqrt{T-t}$ where $Z \sim N(0,1)$, the expectation is expressed as an integral with respect to the standard normal density; calculating it, we get the Black-Scholes formula (7.3).

2.2 Further developments

The discussion above shows that any option payoff $h(\cdot)$ is priced by the discounted expectation formula (7.13). By the martingale representation theorem for Brownian motion we always have

$$e^{-rT}h(S_T) = e^{-rT}\mathbb{E}_{\mathbb{Q}}[h(S_T)] + \int_0^T \psi(t)\,d\check{W}_t$$

for some integrand ψ. Then, from (7.11) the hedge strategy Δ is given by $\Delta_t = \psi_t/\sigma\tilde{S}_t$.

Over a 20-year period following the appearance of the Black-Scholes formula the exact relationships between absence of arbitrage, the existence of risk-neutral measures and the ability to replicate option payoffs were worked out, see Davis and Etheridge (2006) for an account of the historical development. The final product (see Delbaen and Schachermayer (2008), or the compressed accounts Schachermayer (2010), Biagini (2010)) are the two *Fundamental Theorems of Asset Pricing (FTAP)* relating to a semimartingale price process S_t on a given probability space $(\Omega, \mathcal{F}, \mathbb{P})$. FTAP-I states that existence of an equivalent local martingale measure is equivalent to the technical no-arbitrage condition NFLVR ("no free lunch with vanishing risk"), while FTAP-II states that there is a martingale representation theorem, i.e., hedging strategies can be constructed as described above, if and only if the equivalent local martingale measure is unique.

Note that fixing the measure $\mathbb{P}$ is equivalent to specifying a "model" for the price process S_t, in that all its finite-dimensional distributions are then determined.

2.3 Implied volatility and market trading

What happens if we attempt to use Black-Scholes delta hedging in real market trading? This question has been considered by several authors, including El Karoui et al. (1998) and Fouque et al. (2000). The Black-Scholes formula has a certain "robustness" property and can be meaningfully applied in cases where the "real" process driving the market is far away from the stylized Black-Scholes model (7.1). Surprisingly, it took some while for this fact to be noticed; it is hard to imagine that the derivatives market could exist at all without some such property, given that the price model (7.1) is known not to be a particularly accurate representation of real financial price series.

The Black-Scholes formula (7.3) (with $t = 0$) has five parameters $C = C(T, K, S_0, r, \sigma)$. The first four are "known": (K, T) are the contract specification while (S_0, r) are observable market data. So we can regard the formula as a mapping $\sigma \mapsto C(\sigma)$, taking values in the interval $[e^{-rT}[F-K]^+, e^{-rT}F)$ where F is the forward price $F = S_0 e^{(r-q)T}$. This map is increasing in σ, so if we are given the price p of an option (determined by market trading) lying within the allowable range, there is a unique *implied volatility* $\hat{\sigma}$ satisfying $p = C(\hat{\sigma})$. If the underlying price process S_t actually was geometric Brownian motion (7.1) then $\hat{\sigma}$ would be the same, and equal to the volatility σ, for call options of all strikes and maturities. Of course, this is never the case in practice, but we can examine what happens if we naïvely apply the Black-Scholes delta-hedge when in reality the underlying process is *not* geometric Brownian motion. Let us assume that the "true" price model, under measure $\mathbb{P}$, is

$$S_t = S_0 + \int_0^t \eta_t S_{t-} \mathrm{d}t + \int_0^t \kappa_t S_{t-} \mathrm{d}W_t, \tag{7.14}$$

in which η_t, κ_t are general $\mathcal{F}_t$-measurable processes–we do not assume they are deterministic or local coefficients of the form $\eta(t, S_t)$ for example. Consider the scenario of selling at time 0 a European call option at implied volatility $\hat{\sigma}$, i.e., for the price $p = \mathcal{C}(T, S_0, K, r, \hat{\sigma})$ and then following a Black-Scholes delta-hedging trading strategy based on constant volatility $\hat{\sigma}$ until the option expires at time T. As usual, we shall denote $C(t,s) = \mathcal{C}(T-t, K, s, r, \hat{\sigma})$, so that the hedge portfolio, with value process X_t, is constructed by holding $\Delta_t := \partial_S C(t, S_{t-})$ units of the risky asset S, and the remainder $\beta_t := \frac{1}{B_t}(X_{t-} - \Delta_t S_{t-})$ units in the riskless asset. This portfolio, initially funded by the option sale (so $X_0 = p$), defines a self-financing trading strategy. Hence the portfolio value process X satisfies the SDE

$$X_t = p + \int_0^t \partial_S C(u, S_{u-}) \eta_u S_{u-} \mathrm{d}u + \int_0^t \partial_S C(u, S_{u-}) \kappa_u S_{u-} \mathrm{d}W_u$$

Now define $Y_t = C(t, S_t)$, so that in particular $Y_0 = p$. Applying the Itô formula gives

$$\begin{aligned} Y_t = p &+ \int_0^t \partial_t C(u, S_{u-}) \mathrm{d}u + \int_0^t \partial_S C(u, S_{u-}) \eta_u S_{u-} \mathrm{d}u \\ &+ \int_0^t \partial_S C(u, S_{u-}) \kappa_u S_{u-} \mathrm{d}W_u + \frac{1}{2} \int_0^t \partial^2_{SS} C(u, S_{u-}) \kappa_u^2 S_{u-}^2 \mathrm{d}u \end{aligned}$$

Thus the "hedging error" process defined by $Z_t := X_t - Y_t$ satisfies the SDE

$$\begin{aligned} Z_t &= \int_0^t r X_u \mathrm{d}u - \int_0^t (r S_{u-} \partial_S C(u, S_{u-}) + \partial_t C(u, S_{u-}) + \frac{1}{2} \kappa_u^2 S_{u-}^2 \partial^2_{SS} C(u, S_{u-})) \mathrm{d}u \\ &= \int_0^t r Z_u \mathrm{d}u + \frac{1}{2} \int_0^t \Gamma(u, S_{u-}) S_{u-}^2 (\hat{\sigma}^2 - \kappa_u^2) \mathrm{d}u \end{aligned} \tag{7.15}$$

where $\Gamma(t, S_t) = \partial^2_{SS} C(t, S_t)$, and the last equality follows from the Black-Scholes PDE. Therefore the final difference between the hedging strategy and the required option payout is given by

$$Z_T = X_T - [S_T - K]^+ = \frac{1}{2}\int_0^T e^{r(T-t)} S_{t-}^2 \Gamma(t, S_{t-})(\hat{\sigma}^2 - \kappa_t^2)\mathrm{d}t \qquad (7.16)$$

Equation (7.16) is a key formula, as it shows that successful hedging is quite possible even under significant model error. It depends entirely on the relationship between the implied volatility $\hat{\sigma}$ and the true "local volatility" κ_t. For a call or put option $\Gamma_t > 0$. If we, as option writers, are lucky and $\hat{\sigma}^2 > \kappa_t^2$ a.s. for all t then the hedging strategy makes a profit *with probability one* even though the true price model is substantially different from the assumed model (7.1). On the other hand if we underestimate the volatility we will consistently make a loss. The magnitude of the profit or loss depends on the option convexity Γ. If Γ is small then the hedging error is small even if the volatility has been grossly mis-estimated.

3 The volatility surface

In a traded option market, prices are of course determined by supply and demand in the market. Standardized contracts are traded and we will consider a market such as the S&P500 index in which call (and put) contracts are defined with exercise times $T_1,\ldots,T_N$ and, for each T_i, a range of strikes $\{K_{ij}, 1 \le j \le M_i\}$. The corresponding market prices and Black-Scholes implied volatilities are denoted p_{ij} and $\hat{\sigma}_{ij}$. The latter define the *volatility surface*, which would be flat (i.e., $\hat{\sigma}_{ij} \equiv \sigma$ for all i,j) if the Black-Scholes model with volatility σ were correct, but in reality exhibits a "smile" or a "smirk". The reader can consult Gatheral (2006) for detailed description. It is of course a matter of primary importance how to use this information to identify suitable models and hedging strategies. The first step is to interpolate and extrapolate the given finite data $\hat{\sigma}_{ij}$ to give a continuous-parameter surface $\{\hat{\sigma}(t,k), t \in [0,T_N), k \ge 0\}$ with corresponding prices $p(t,k)$. This surface contains a surprising amount of information, as detailed next.

NOTE: Throughout the remainder of this chapter we shall assume, purely for ease of exposition, that interest rates and dividend yield are zero, and hence that price processes are martingales under a risk-neutral measure. This assumption is harmless since it is equivalent to assuming that we are working with forward prices rather than spot prices, and in our framework (no interest rate volatility) there is a one-to-one relationship between forward and spot.

3.1 The Breeden-Litzenberger formula (Breeden and Litzenberger 1978)

As mentioned above, we set interest rates and dividend yields to zero here (or, equivalently, work in terms of forward prices). Using the risk-neutral representation (7.13) of prices we have

$$p(t,k) = \int_{\mathbb{R}^+} [s-k]^+ \mu_t(ds) = \int_{(k,\infty)} (s-k)\mu_t(ds), \tag{7.17}$$

where μ_t is the distribution function of S_t. Hence

$$\frac{\partial}{\partial k} p(t,k) = -\int_{(k,\infty)} \mu_t(ds) = \mu_t(k+) - 1, \tag{7.18}$$

and if μ_t has a density function ϕ_t then

$$\frac{\partial^2}{\partial k^2} p(t,k) = \phi_t(k). \tag{7.19}$$

The key message is that the strip of option prices with maturity t determines the risk-neutral distribution of S_t.

3.2 The Dupire equation (Dupire 1994)

Suppose the price process S_t satisfies a *local volatility model*

$$dS_t = S_t \sigma(t, S_t) dW_t, \quad t \in [0, T_N] \tag{7.20}$$

Assume the solution of this equation has density $\phi_t(s)$. Then ϕ satisfies the Kolmogorov forward equation

$$\frac{\partial}{\partial t}\phi_t(s) = \frac{1}{2}\frac{\partial^2}{\partial s^2}(s^2\sigma^2(t,s)\phi_t(s)).$$

From (7.17)

$$\begin{aligned} \frac{\partial p}{\partial t} &= \int_k^\infty \frac{\partial}{\partial t}\phi_t(s)(s-k)ds \\ &= \frac{1}{2}\int_k^\infty \frac{\partial^2}{\partial s^2}(s^2\sigma^2(t,s)\phi_t(s))(s-k)\,ds. \end{aligned} \tag{7.21}$$

Integrating by parts in (7.21) gives

$$\begin{aligned} \int_k^\infty \frac{\partial^2}{\partial s^2}(s^2\sigma^2(t,s)\phi_t(s))(s-k)ds &= \int_k^\infty -\frac{\partial}{\partial s}(s^2\sigma^2\phi)\frac{\partial}{\partial s}(s-k)ds \\ &= -\int_k^\infty \frac{\partial}{\partial s}(s^2\sigma^2\phi)ds \\ &= k^2\sigma^2(t,k)\phi_t(k), \end{aligned} \tag{7.22}$$

since all the boundary terms are zero. Hence (7.21) and (7.22) give us the *Dupire equation*

$$\frac{\partial}{\partial t}p(t,k) = \frac{1}{2}k^2\sigma^2(t,k)\frac{\partial^2}{\partial k^2}p(t,k).$$

Rearranging this gives

$$\sigma(t,k) = \sqrt{\frac{2\partial p/\partial t}{k^2 \partial^2 p/\partial k^2}}. \tag{7.23}$$

Remarkably, the volatility surface determines the model (7.20). This is certainly a striking result, but it has some limitations. Although the model is produced directly from the data, there is an *a priori* assumption that the model takes the form (7.20). There is no simple test that can be applied to the data to check whether a model in this class is appropriate. The model the formula produces is heavily dependent on the interpolation/extrapolation algorithm used to generate the whole surface from discrete data. Finally, there is no time-consistency: if we compute the volatility function $\sigma(\cdot,\cdot)$ from today's volatility surface and then recompute tomorrow using new data we will invariably get a different result. In fact the main use of the Dupire formula is to produce valuations for other options consistent with the prices of the traded options, and for that purpose it can be very effective.

3.3 Stochastic volatility models

In view of the limitations of local volatility models of the form (7.20) there has in recent years been a huge amount of work devoted to stochastic volatility models in which the function σ depends on additional random factors, not just on the price process itself. A typical example is the Heston model

$$dS_t = \mu(t) S_t dt + \sqrt{v_t} S_t dW_t,$$

$$dv_t = \lambda(\theta - v_t)dt + \eta\sqrt{v_t}\, dZ_t,$$

where W,Z are correlated Brownian motions. The second equation is an autonomous "CIR"-type equation generating the local variance v_t. One consequence of the extra randomness is that the model is no longer complete if S_t and the riskless asset are the only traded assets. We need one more traded asset to complete the market. Without it, there are multiple risk-neutral measures and options cannot be perfectly replicated, by the second FTAP.

We do not discuss stochastic volatility models further in this chapter, referring the reader to Gatheral (2006) for an authoritative account. Rather than following up properties of specific models, our intention here is to investigate what more the price data can tell us about pricing without making an *a priori* model choice.

4 Known Marginals

Our starting point in this section is the finite set of options data described at the beginning of Section 3. Suppose we agree to interpolate/extrapolate, for each exercise time T_i, the implied volatilities $\{\hat{\sigma}_{ij}, 1 \le j \le M_i\}$ and hence to determine, via the Breeden-Litzenberger formula (7.18), the risk-neutral marginal distribution μ_i of S_{T_i}. Any risk-neutral model for the price process $\{S_t, t \in [0, T_N]\}$ must be a martingale and must respect these one-dimensional distributions. The conditions under which, for a given set of measures μ_i, there exists a martingale measure $\mathbb{Q}$ such that $S_{T_i} \sim \mu_i$ for each i are a special case of a more general result of Strassen (1965). They are that the measures μ_i must have the same finite mean and be *increasing in convex order*, i.e., for any convex function ψ the integrals $\int_{\mathbb{R}+} \psi(s)\mu_i(ds)$ must be increasing in i. Let $\mathcal{Q}$ be the set of martingale measures when these conditions are satisfied. If they fail then by the FTAP there is an arbitrage opportunity since there is no equivalent martingale measure.

Suppose we now have another contract whose exercise value at time T_N is $\Phi(S_t, 0 \le t \le T_N)$, a possibly "path-dependent" function. Then the range of no-arbitrage valuations of Φ is $R = \{\mathbb{E}_{\mathbb{Q}}[\Phi(S(\cdot))] : \mathbb{Q} \in \mathcal{Q}\}$. Since $\mathcal{Q}$ is a convex set, R is an interval. To determine R and give some information about hedging strategies, a variety of techniques have been introduced.

4.1 Static replication

Consider first a simple European option maturing at time T_i with exercise value $\Phi(S) = f(S_{T_i})$ where f is a convex function. Then $f : \mathbb{R}^+ \to \mathbb{R}$ has a right derivative $f'_+(x)$ which is an increasing function of $x \in \mathbb{R}^+$, so the recipe $f''(a,b] = f'_+(b) - f'_+(a)$ defines a positive measure $f''(dx)$ on $\mathcal{B}(\mathbb{R}^+)$, equal to $f''(x)dx$ if f is C^2. We then have the second order exact Taylor formula

$$f(x) = f(x_0) + f'_+(x_0)(x - x_0) + \int_0^{x_0} (y - x)^+ f''(dy) + \int_{x_0}^{\infty} (x - y)^+ f''(dy). \tag{7.24}$$

Evaluating this at $x = S_{T_i}$, taking the risk-neutral expectation and using the Fubini theorem, gives

$$\begin{aligned} \mathbb{E}_{\mathbb{Q}}[f(S_{T_i})] &= f(S_0) + \int_0^{S_0} \mathbb{E}_{\mathbb{Q}}[y - S_{T_i}]^+ f''(dy) + \int_{S_0}^{\infty} \mathbb{E}_{\mathbb{Q}}[S_{T_i} - y]^+ f''(dy) \\ &= f(S_0) + \int_0^{S_0} \mathcal{P}(y) f''(dy) + \int_{S_0}^{\infty} \mathcal{C}(y) f''(dy) \end{aligned} \tag{7.25}$$

where $\mathcal{C}(y), \mathcal{P}(y)$ are the call and put prices at strike y. Since $\mathcal{C}(y)$ is the strip of call option prices obtained from the market volatility surface, and the values $\mathcal{P}(y)$ are determined by put–call parity, we see that (7.25) gives a model-free evaluation of the new option price. We don't even need the Breeden-Litzenberger formula for this. Further, by plugging $x = S_{T_i}$ into (7.24) we obtain

$$f(S_{T_i}) = f(S_0) + f'_+(S_0)(S_{T_i} - S_0) + \int_0^{S_0} [y - S_{T_i}]^+ f''(dy) + \int_{S_0}^{\infty} [S_{T_i} - y]^+ f''(dy). \tag{7.26}$$

This is *static replication*: the option payoff is exactly replicated by a portfolio consisting of cash, a static holding in the underlying asset and a weighted sum of put and call options.

4.2 Skorokhod embedding

We refer the reader to Hobson (2010) and Obłój (2004 2010) for comprehensive expositions of this topic. Skorokhod embedding refers to the following question: given a probability measure μ on $\mathbb{R}$ and a Brownian motion B_t with $B_0 = 0$, can we find a stopping time τ such that $B_\tau \sim \mu$ (i.e., the distribution of B_τ is μ)? The answer is yes: let F be the distribution function for μ. Then since $B_1 \sim N(0,1)$, $F^{-1}(N(B_1)) \sim \mu$ where N is given by (7.4). Hence the stopping time $\tau = \inf\{t \geq 2 : B_t = F^{-1}(N(B_1))\}$ satisfies $B_\tau \sim \mu$. However, this is not satisfactory since among other things $\mathbb{E}[\tau] = \infty$. For applications, we need the process $B_{t\wedge\tau}$ to be a uniformly integrable martingale (so that in particular $\mathbb{E}[B_\tau] = 0$). Decades of research on this topic have produced a cornucopia of "good" solutions, for details of which the reader can consult the references.

The connection with finance is as follows. Suppose we have traded options maturing at a single time T and we back out the corresponding risk-neutral distribution μ_T à la Breeden-Litzenberger. Let τ be a solution of the Skorokhod problem for μ_T and define $S_t = B_{\tau\wedge(t/(T-t))}$. Then S_t is a martingale such that $S_T \sim \mu_T$. We have to make the obvious modifications to the embedding procedure so that S_0 is equal to the initial asset price. Then we have created a model in a risk-neutral measure $\mathbb{Q}_\tau$, that correctly prices all the traded assets in the market. Different models are created by different solutions of the Skorokhod problem.

Suppose we now have another contract with possibly path dependent payoff $\Phi(S)$ at time T. We can seek to establish bounds on the value of this contract by calculating $\inf\{\mathbb{E}_{\mathbb{Q}_\tau}[\Phi(S)]: \tau \in \mathfrak{T}\}$ and $\sup\{\mathbb{E}_{\mathbb{Q}_\tau}[\Phi(S)] : \tau \in \mathfrak{T}\}$, where $\mathfrak{T}$ indexes some class of solutions to the Skorokhod problem. This generally has to be done on a case-by-case basis, see for example Hobson (1998), Brown, Hobson, and Rogers (2001), Cox and Obłój (2011ab), where the associated hedging strategies are also studied. The technique works best in case where the payoff $\Phi(S)$ is invariant under time change; for example if we have a one-touch option where $\Phi(S)$ is equal to 1 if $\sup_{0<t\leq T} S_t \geq b$ for some barrier $b > S_0$ and 0 otherwise. Obviously, this is the same event as $\sup_{0<t\leq\infty} B_{t\wedge\tau} \geq b$, so we don't have to consider the slightly awkward mapping from B to S. Methods based on stochastic control to handle a wider range of problems are developed by Galichon, Henry-Labordère, and Touzi (2014).

4.3 Optimal transport

This problem goes back to the 18th century. We first outline the original formulation by Monge (1781) in terms of "déblais" and "remblais" in a one-dimensional version. The physical interpretation is that we are given a pile of soil or rubble (the déblais), with mass density f_1, which we wish to transport to an embankment (the remblais), with mass density f_2 (the total mass must of course be the same, so we can normalize to total mass = 1). Mass at x is transported to $y = g(x)$ where g is a smooth 1–1 function; the condition under which the resulting mass density will indeed be f_2 is

$$f_1(x) = f_2(g(x))g'(x). \tag{7.27}$$

As each particle of soil moves a distance $|x - g(x)|$, the total work involved is

$$I(g) = \int_{\mathbb{R}} |x - g(x)| f_1(x) dx.$$

The problem is to find a rearrangement function g that requires the least work.

This problem has been the subject of extensive study (cf. Evans and Gangbo 1999, Villani 2009) ever since its introduction and is awkwardly non-linear because of the constraint (7.27). Centuries later, Kantorovich (1942) formulated a much simpler version in which, instead of requiring functional dependence between x and y we specify a joint distribution of (x, y) such that the marginals are f_1 and f_2. Generalizing a little from the setting above, our problem is now to minimize a work function

$$I(\mu) = \int_{\mathbb{R}^2} c(x, y)\mu(dx, dy) \tag{7.28}$$

over all probability measures μ on $\mathbb{R}^2$ with given marginals μ_1, μ_2, a constraint we can state as

$$\int_{\mathbb{R}^2} (u(x) + v(y))\mu(dx, dy) = \int_{\mathbb{R}} u(x)\mu_1(dx) + \int_{\mathbb{R}} v(x)\mu_2(dx), \quad u, v \in C_b(\mathbb{R}). \tag{7.29}$$

This is now an infinite-dimensional linear programming (LP) problem: minimize the linear function $I(g)$ of (7.28) subject to linear constraints (7.29). As in the standard finite-dimensional case, there is a dual linear program

$$\begin{aligned} &\text{Maximize} \quad J(u, v) = \int_{\mathbb{R}} u(x)\mu_1(dx) + \int_{\mathbb{R}} v(y)\mu_2(dy) \\ &\text{subject to} \quad u(x) + v(y) \le c(x, y),\ (x, y) \in \mathbb{R}^2. \end{aligned} \tag{7.30}$$

Clearly $J(u, v) \le I(\mu)$ if (u, v) and μ satisfy the constraints, so we have "weak duality". In fact strong duality, i.e., $\sup_{u,v} J(u, v) = \inf_{\mu} I(\mu)$, holds under minimal technical conditions. The theory extends without difficulty to multi-period cases where μ is a measure on $\mathbb{R}^n$ and (u, v) are replaced by $(u_1, \ldots, u_n)$.

The connection with finance is easy to see. Suppose we have an options market with contracts maturing at times $T_1, \ldots, T_n$. Then given prices (or implied volatilities) for sufficiently many options we can use Breeden-Litzenberger to determine the marginal distributions μ_i of the underlying price S_{T_i} for $i = 1, \ldots, n$. Recall that these are the marginal distributions in any risk-neutral measure. Thus the class $\mathfrak{M}$ of *consistent* risk-neutral measures $\mathbb{Q}$ consists of all *martingale* measures on $\mathbb{R}^n$ having the right marginals at all T_i. The martingale property is an additional feature not present in general optimal transport, but in fact is just another linear constraint. Given an option payoff $\Phi(S_{T_1},\ldots,S_{T_n})$, we can define

$$I_\Phi(\mathbb{Q}) = \mathbb{E}_\mathbb{Q}[\Phi(S_{T_1},\ldots,S_{T_n})] = \int_{\mathbb{R}^n} \Phi(\mathbf{x})\mathbb{Q}(d\mathbf{x}), \quad \mathbb{Q} \in \mathfrak{M},$$

and one estimate of the option value could be

$$\overline{p}_\Phi = \inf_{\mathbb{Q}\in\mathfrak{M}} I_\Phi(\mathbb{Q}). \tag{7.31}$$

A more direct estimate is based on the idea of sub-replication. Suppose we form a portfolio of traded assets whose value at time 0 is p. If the portfolio value at T_n is $X(S_{T_1},\ldots,S_{T_n})$ and $X(\mathbf{x}) \le \Phi(\mathbf{x})$ for all $\mathbf{x} \in (\mathbb{R}^+)^n$ then there is an obvious arbitrage if the payoff Φ can be bought for less than p. So our best lower estimate for the value of Φ is the value of the most expensive sub-replicating portfolio. Recall from Section 4.1 that an arbitrary (subject to integrability conditions) option payoff $f(S_{T_i})$ can be replicated exactly using the representation (7.26), with value (7.25) completely specified by the volatility surface at exercise time T_i. We can also, at zero cost, trade in the underlying asset at all intermediate times. Thus candidates for sub-hedging are "semi-static" portfolios of the form

$$X(x_1,\ldots,,x_n) = \sum_{i=1}^{n} u_i(x_i) + \sum_{i=0}^{n-1} \Delta_i(x_1,\ldots,x_i)(x_{i+1} - x_i). \tag{7.32}$$

The time-0 value of X is just

$$p(X) = \sum_{i=1}^{n} \int_{\mathbb{R}^+} u_i(x)\mu_i(dx) = \mathbb{E}_\mathbb{Q}[X] \quad \forall \mathbb{Q} \in \mathfrak{M}.$$

The sub-replication problem is to maximize $p(X)$ over all $(u_1,\ldots,u_n,\Delta)$ such that $X(\mathbf{x}) \le \Phi(\mathbf{x})$ for all $\mathbf{x} \in (\mathbb{R}^+)^n$. Denote by $\underline{p}_\Phi$ the supremum of such $p(X)$; then it is clear that $\underline{p}_\Phi \le \overline{p}_\Phi$. In fact, we have strong duality, i.e., $\underline{p}_\Phi = \overline{p}_\Phi$, under quite general conditions, and there exists a $\mathbb{Q}$ that attains the infimum in (7.31); these results are due to Beiglböck, Henry-Labordère, and Penkner (2013). Their main condition is that Φ be a lower semi-continuous function such that

$$\Phi(\mathbf{x}) \ge -\kappa(1 + |x_1| + \cdots + |x_n|), \quad \mathbf{x} \in (\mathbb{R}^+)^n,$$

a condition satisfied in most potential applications.

If we were to omit the martingale condition, we would get a lower value for the infimum in (7.31). We could then look at the standard Kantorovich dual problem (7.30), which is maximizing over X as in (7.32) but *without the final re-hedge term* – so the supremum is lower. The results of Beiglböck et al. (2013) show that the magnitudes of these effects are exactly the same.

In conventional mathematical finance, with a fixed model, it is a well-established principle (see Föllmer and Schied 2011, Theorem 5.29) that the range of arbitrage-free prices of a contingent claim is equal to the set of conditional expectations under all risk-neutral measures and the upper and lower bounds of this set are given by the costs of the cheapest super-replicating and most expensive sub-replicating strategies respectively. It is striking that this principle extends to the model-free setting we have just described.

So far, only a few cases have been examined in detail (Hobson and Neuberger 2012, Hobson and Klimmek 2015) and general computational methods have not been investigated. This is work in progress.

5 Option data without interpolation

In the previous section we assumed that the entire volatility surface was known at a finite number of exercise times. This gives elegant solutions since then the marginal distributions are determined by Breeden and Litzenberger (1978) and the set of models consistent with this data coincides with the set of martingales with these marginals. The basic assumption is however only credible in markets in which a large number of strikes are traded for each exercise time, which in practice means the big index markets such as the S&P500 or the FTSE100. In other circumstances we have to deal directly with the finite set of option data we actually have, i.e., the current market prices p_{ij} for European call options with exercise time T_i and strike K_{ij}, $i=1,\ldots,N, j=1,\ldots,M_i$. As above we assume "frictionless trading," i.e., no bid–ask spread, although this could, with some pain, be included in some of the questions below.

The first question, studied by Davis and Hobson (2007), is whether arbitrage is somehow built in to the prices p_{ij}. It turns out that there are three cases: (i) the prices are consistent with absence of arbitrage, (ii) there is a model-independent arbitrage, and (iii) there is a model-dependent arbitrage. Case (i) means that we can construct a joint distribution μ for $S=(S_0,S_{T_1},\ldots,S_{T_N})$ such that S_0 is equal to today's price with probability 1, the process S is a martingale, and for each i,j

$$p_{ij}=\int_{(\mathbb{R}^+)^{N+1}}[s_i-K_{ij}]^+\mu(ds), \tag{7.33}$$

that is, the prices are expectations in a risk-neutral measure μ. In Case (ii), we can trade at time 0 in such a way that we make an immediate profit and are left with a portfolio whose value is always non-negative; this is an arbitrage opportunity. In Case (iii) we know we are not in Case (i), i.e., that a representation as in (7.33) is impossible, but without further information we cannot tell what trading strategy will realize risk-free profit. For a single time T_i the conditions for being in Case (i) are illustrated in Figure 7.1: the linear interpolant of the prices p_{ij} plotted against strikes $K_{ij}, j=1,\ldots,M_i$ is a strictly decreasing convex function with slope ≥ -1 at the origin (we include S_0 as the price of an option with strike 0). Case (iii) arises when these conditions are met except that the

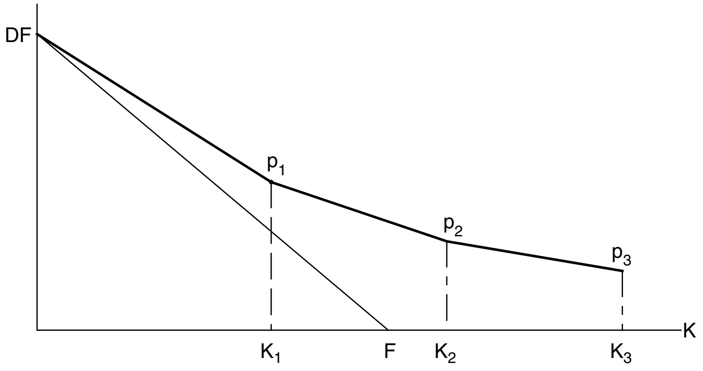


Figure 7.1 Call prices consistent with absence of arbitrage

interpolant is not *strictly* decreasing; otherwise we are in Case (ii). To explain Case (iii): this arises when we have two options with different strikes but the same strictly positive price: $p_{ij} = p_{i(j+1)}$. A moment's thought shows that there is no possibility of representing these prices in terms of a risk-free measure as in (7.33), but what strategy realizes the arbitrage? The answer depends on the range of prices that have positive probability. If there is some chance that the price will rise above K_{ij}, sell the option with strike $K_{i(j+1)}$ and buy K_{ij}. This costs nothing and creates a spread option whose value is always non-negative and may be positive: an arbitrage opportunity. If, on the other hand, the price will never exceed K_{ij} then the arbitrage strategy is to sell either option; it will never be exercised. Without knowing which condition applies we cannot construct a strategy that is guaranteed to realize arbitrage.

When we consider the options at all times simultaneously, the conditions for consistency with absence of arbitrage are easily stated when interest rate and dividend yields are zero (or equivalently, we work in normalized prices), $M_i = M_1$ for all i and $K_{ij} = K_{1j}$ for all i,j, i.e., we have the same set of strikes at each exercise time. Then the prices p_{ij} are consistent with absence of arbitrage if (a) the conditions above are met at each T_i, and (b) all "calendar spreads" with exercise value $[S_{T_{i+1}} - K_{(i+1)j}]^+ - [S_{T_i} - K_{ij}]^+$ have non-negative price. When the strikes are not all the same at different times the condition is similar in spirit but is somewhat complicated to state and involves all the strikes jointly (see Davis and Hobson 2007, Theorem 4.2).

Given a set of options prices satisfying the consistency conditions, we can determine arbitrage bounds on the prices of other options. Consider first, as in Davis, Obłój, and Raval (2014), the situation where all options are put options maturing at a single time T with strikes and prices $(k_i,p_i), i = 1,\dots,m$, and we wish to determine a lower bound on the value of a convex payoff $\Phi(S_T)$. At time 0 we can form a static portfolio of cash, the underlying asset and the traded options. The value of such a portfolio at time T when $S_T = s$ will be $\sum_{i=1}^{n} y_i a_i(s)$ where $n = m + 2$ and

$$\begin{aligned} a_1(s) &= 1 \quad \text{(cash)} \\ a_2(s) &= s \quad \text{(underlying)} \\ a_{i+2}(s) &= [k_i - s]^+, \quad i = 1, \dots, m \quad \text{(options)} \end{aligned}$$

and y_i is the number of units of asset i in the portfolio (we normalize to $S_0 = 1$). A lower bound for the value of Φ is the value of the *most expensive sub-replicating portfolio*, and this is the solution of the *semi-infinite linear program (LP)*

$$\mathfrak{P}: \quad \sup_{\mathbf{y} \in \mathbb{R}^n} \mathbf{y}'\mathbf{b} \quad \text{subject to} \quad \mathbf{y}'\mathbf{a}(s) \le \Phi(s) \; \forall s \in \mathbb{R}^+,$$

where $\mathbf{y}$ is the vector of portfolio weights y_i, $\mathbf{b}$ is the price vector $b_1 = b_2 = 1, b_i = p_{i-2}$, $i = 3, \dots, n$ and $\mathbf{a}$ is the vector with components a_i as above. For this problem, we can apply the Karlin-Isii duality theorem of semi-infinite LP (Karlin and Studden 1966) to conclude that the value of the primal problem $\mathfrak{P}$ is equal to the value of the dual LP

$$\mathfrak{D}: \quad \inf_{\mu \in \mathbb{M}} \int_{\mathbb{R}+} \Phi(s)\mu(ds) \quad \text{subject to} \quad \int_{\mathbb{R}+} \mathbf{a}(s)\mu(ds) = \mathbf{b},$$

where $\mathbb{M}$ is the set of positive Borel measures such that each a_i is integrable. As long as the option prices satisfy the Davis-Hobson conditions there is no duality gap, and we always have existence in $\mathfrak{P}$; existence may however fail for $\mathfrak{D}$. When existence holds there is a very clear interpretation of the optimal dual measure. The optimal sub-replicating portfolio $\mathbf{y}'\mathbf{a}(s)$ is piecewise linear and is tangential to Φ at a finite number of points $s_1, \dots, s_k$; the optimal measure for $\mathfrak{D}$ takes the form $\mu(ds) = \sum_1^k \alpha_i \delta_{s_i}(ds)$, a sum of Dirac measures on the points s_i. The sub-replication constraint in $\mathfrak{P}$ is "slack" at all $s \notin \{s_1, \dots, s_k\}$, so this characterization is analogous to the "complementary slackness" property of finite-dimensional LPs.

The same problem but including traded options maturing at $T' < T$ in addition to those maturing at T is more complicated, because then it is natural to consider re-hedging by trading in the underlying at the intermediate time T'. The time-T value of the resulting "semi-static" portfolio will then take the form (with $s' = S_{T'}, s = S_T$)

$$f(s', s) = y_1 + y_2 s + \sum y_i'[k_i' - s']^+ + \sum y_i[k_i - s']^+ + \Delta(s')(s - s'), \qquad (7.34)$$

where Δ defines our re-hedge strategy at T'. Since Δ is a function of the continuous parameter s' we now have an infinite-dimensional decision variable and hence a doubly-infinite LP, requiring more functional-analytic technique.

A recent paper by Acciaio et al. (2015) has established an FTAP in this setting. They consider an n-stage discrete-time model so that a price path $\mathbf{s}$ is identified with an element $x \in \Omega = (\mathbb{R}^+)^n$. Options with general payoffs $\{\varphi_i(x) : i \in I\}$ are available at (without loss of generality) zero cost, and a dynamic trading strategy in the underlying produces a gain from trade

$$(\Delta \cdot x)_n = \sum_{k=0}^{n-1} \Delta_k(x_1,\ldots,x_k)(x_{k+1} - x_k),$$

generalizing the single-period re-hedge in (7.34). There is a *model-independent arbitrage* if $\sum_{j=1}^{N} a_j \varphi_{i_j}(x) + (\Delta \cdot x)_n > 0$ for some choice of indices $i_1,\ldots,i_N$, trading strategy Δ and constants $a_1,\ldots,a_N$. (Recall that by assumption the cost of entering this trade is zero.) Now let $\mathcal{M}$ be the set of probability measures $\mathbb{Q}$ on Ω such that (i) the "process" x is a $\mathbb{Q}$-martingale and (ii) $\int \varphi_i d\mathbb{Q} \le 0$ for all $i \in I$.

The conditions required for the FTAP are as follows. First, there is a convex super-linear function g such that $\varphi_{i^*}(x) = g(x_n)$ for some $i^* \in I$. Second, with $m(x) = \sum_{j=1}^{n} g(x_j)$, we have for all $i \in I$

$$\lim_{||x|| \to \infty} \frac{\varphi_i(x)^+}{m(x)} < \infty, \qquad \lim_{||x|| \to \infty} \frac{\varphi_i(x)^-}{m(x)} = 0.$$

This second condition rules out model-dependent arbitrage of the sort encountered in Davis and Hobson (2007). The FTAP then states that under the above two conditions the following are equivalent:

(i) There is no model-independent arbitrage.
(ii) $\mathcal{M} \neq \emptyset$.

Some analogous results for models including bid–ask spreads have been obtained by Dolinsky and Soner (2014). Applications to the valuation of variance swaps are considered in Davis et al. (2014) where it turns out that valuations implied by maximal sub-replication are surprisingly close to actual market prices.

6 Concluding Remarks

It could be argued that the methods and results described in this chapter take a rather extreme point of view. In order to free ourselves from the tyranny of a specific model, represented by a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, we have thrown away *all* models, retaining only a database of current market prices, and allowed ourselves only to make statements about price bounds that do not imply arbitrage of a very simple kind. How useful these statements will turn out to be will certainly depend on the problem. Some option payoffs are well approximated by static portfolios plus some trading in the underlying, in which case we expect the bounds to be tight, while in other cases the option payoff is very different to anything constructed from current puts and calls and the bounds will be unusably large. So far the repertoire of solved cases in this area is too small to enable us to go beyond this general comment. However, the endeavor seems to this author to be worthwhile, in that we discover just how much is implied by minimal assumptions, and that much turns out to be rather more than one might expect. It also focuses attention on the deficiencies of another extreme point of view: that one specific model is correct!

As usual, the ultimate answer may lie in some form of compromise, in which some extra assumptions of an economic character are brought in to tighten the bounds. This has been done with some success in the "fixed model" arena in the literature on "good deal bounds" (Cochrane and Saá-Requejo 2000, Björk and Slinko 2006). Here putative prices are rejected if investing in assets at these prices together with other market-traded assets would lead to portfolios with unrealistically high Sharpe ratios. Perhaps some such method could be extended to the model-free arena.

References

Acciaio, B., M. Beiglböck, F. Penkner, and W. Schachermayer (2015). A model-free version of the fundamental theorem of asset pricing and the super-replication theorem. *Mathematical Finance*. (Published online December 2013).

Bachelier, L. (1900). Théorie de la spéculation. *Annales Scientifiques de l'Ecole Normale Supérieure 17*, 21–86.

Beiglböck, M., H. Henry-Labordère, and F. Penkner (2013). Model-independent bounds for option prices: A mass transport approach. *Finance and Stochastics 17*, 477–501.

Biagini, F. (2010). Second fundamental theorem of asset pricing. In Cont (2010), pp. 1623–1628.

Björk, T. and I. Slinko (2006). Towards a general theory of good-deal bounds. *Review of Finance 10*, 221–260.

Black, F. and M. Scholes (1973). The pricing of options and corporate liabilities. *The Journal of Political Economy 81*(3), 637–654.

Breeden, D. and R. Litzenberger (1978). Prices of state-contingent claims implicit in option prices. *The Journal of Business 51*, 621–651.

Brown, H., D. Hobson, and L. C. G. Rogers (2001). Robust hedging of barrier options. *Mathematical Finance 11*, 285–314.

Cochrane, J. and J. Saá-Requejo (2000). Beyond arbitrage: Good deal price bounds in incomplete markets. *Journal of Political Economy 108*, 79–119.

Cont, R. (Ed.) (2010). *Encyclopedia of Quantitative Finance*. John Wiley & Sons Inc.

Cox, A. M. G. and J. Obłój (2011a). Robust hedging of double no-touch barrier options. *Finance and Stochastics, 15*, 573–605.

Cox, A. M. G. and J. Obłój (2011b). Robust hedging of double touch barrier options. *SIAM Journal on Financial Mathematics 2*, 141–182.

Davis, M. H. A. (2010). The Black-Scholes formula. In Cont (2010), pp. 199–207.

Davis, M.H.A. (2016). A Beaufort scale of predictability. In M. Podolskij, R. Stelzer, S. Thorbjrnsen, and A. Veraat (Eds), *The Fascination of Probability, Statistics and their Applications*. Springer.

Davis, M. H. A. and A. Etheridge (2006). *Louis Bachelier's Theory of Speculation: The Origins of Modern Finance*. Princeton University Press.

Davis, M. H. A. and D. Hobson (2007). The range of traded option prices. *Mathematical Finance 17*, 1–14.

Davis, M. H. A., J. Obłój, and V. Raval (2014). Arbitrage bounds for weighted variance swap prices. *Mathematical Finance 24*, 821–854.

Delbaen, F. and W. Schachermayer (2008). *The Mathematics of Arbitrage*. Springer.

Dolinsky, Y. and H. M. Soner (2014). Robust hedging with proportional transaction costs. *Finance and Stochastics 18*, 327–347.

Dupire, B. (1994). Pricing with a smile. *Risk 7*, 18–20.

El Karoui, N., M. Jeanblanc-Picqué, and S. E. Shreve (1998). Robustness of the Black and Scholes formula. *Mathematical Finance 8*, 93–126.

Evans, L. C. and W. Gangbo (1999). Differential methods for the the Monge-Kantorovich mass transfer problems. *Memoirs of AMS*, no. 653, vol. 137.

Föllmer, H. and A. Schied (2011). *Stochastic Finance: An Introduction in Discrete Time*. De Gruyter.

Fouque, J.-P., G. Papanicolaou, and K. R. Sircar (2000). *Derivatives in Financial Markets with Stochastic Volatility*. Cambridge University Press.

Galichon, A., H. Henry-Labordère, and N. Touzi (2014). A stochastic control approach to no-arbitrage bounds given marginals, with an application to lookback options. *Annals of Applied Probability 24*, 312–336.

Gatheral, J. (2006). *The Volatility Surface: A Practitioner's Guide*. New York: Wiley.

Glasserman, P. (2003). *Monte Carlo Methods in Financial Engineering*. Springer.

Hilber, N., O. Reichmann, and C. Schwab (2013). *Computational Methods for Quantitative Finance: Finite Element Methods for Derivative Pricing*. Springer.

Hobson, D. (2010). The Skorokhod embedding problem and model-independent bounds for option prices. In R. Carmona, E. Cinlar, E. Ekeland, E. Jouini, J. Scheinkman, and N. Touzi (eds), *Princeton Lectures on Mathematical Finance 2010*, Volume 2003 of *Lecture Notes in Math*. Springer.

Hobson, D. and M. Klimmek (2015). Robust price bounds for forward-starting straddles. *Finance and Stochastics 19*, 189–214.

Hobson, D. and A. Neuberger (2012). Robust bounds for forward start options. *Mathematical Finance 22*, 31–56.

Hobson, D. G. (1998). Robust hedging of the lookback option. *Finance and Stochastics 2*(4), 329–347.

Hull, J. C. (2011). *Options, Futures and Other Derivatives* (8th ed.). Pearson Education.

Kantorovich, L. V. (1942). On the transfer of masses. *Dokl. Akad. Nauk. SSSR 37*, 227–229.

Karlin, S. and W. Studden (1966). *Tchebycheff Systems, with Applications in Analysis and Statistics*. Wiley Interscience.

Monge, G. (1781). Mémoire sur la théorie des déblais et des remblais. *Histoire de l'Académie royale des sciences avec les mémoires de mathématique et de physique tirés des registres de cette Académie*, 666–705.

Obłój, J. (2004). The Skorokhod embedding problem and its offspring. *Probab. Surveys 1*, 321–390.

Obłój, J. (2010). The Skorokhod embedding problem. In Cont (2010), pp. 1653–1657.

Samuelson, P. (1965). Rational theory of warrant pricing. *Industrial Management Review 6*, 13–39.

Schachermayer, W. (2010). The fundamental theorem of asset pricing. In Cont (2010), pp. 792–801.

Strassen, V. (1965). The existence of probability measures with given marginals. *Ann. Math. Statist. 36*, 423–439.

Villani, C. (2009). *Optimal Transport, Old and New*. Springer.

8
The Private Information Price of Risk

Jérôme Detemple and Marcel Rindisbacher

1 Introduction

The Private Information Price of Risk (PIPR) represents the incremental price of risk assessed when private information becomes available. The PIPR plays a prominent role in models with private information. It determines the perception of risk for the recipient of a private information signal. It lies at the heart of the optimal consumption-portfolio policies of such an informed agent. It drives the return performance of an informed fund manager. It is an essential component of the welfare gains derived by investors in professionally managed funds.

This chapter seeks to describe the concept and provide perspective on some of its applications. The approach adopted focuses first on discrete time, then on continuous time models. While the concepts, results and intuitions are the same in both types of settings, they are particularly transparent in the discrete time framework.

In traditional economies with public information, the market price of risk (MPR), also known as the Sharpe ratio, is the fundamental quantity that identifies the reward associated with a fundamental source of risk.[1] The MPR is the expected excess return of an asset normalized by the standard deviation of the return. It represents the risk premium per unit risk, thus captures the reward associated with the underlying risk, i.e., the return innovation. When private information becomes available, the perception of risk changes. An innovation (with zero mean) under public information, may no longer have zero mean in light of the information collected. The change in the innovation mean is the PIPR. It represents the incremental risk premium per unit risk in light of the private signal.

The PIPR modifies the risk-reward trade-off embedded in an asset return. Under private information, the Sharpe ratio of an asset has two components, one associated with public information, the MPR, and one related to private information, the PIPR. The presence of the second component modifies the return characteristics of the asset. It therefore affects the opportunity set of the informed investor, relative an uninformed counterpart, and the eventual payoffs that can be achieved. Optimal choices reflect these revisions in the return properties.

The terminology "PIPR" is introduced in Detemple and Rindisbacher (2013). They use the notion to analyze the returns generated by an informed portfolio manager and examine the implications for timing regressions. The quantity also appears in various studies in mathematical finance dealing with insider trading. It is usually introduced as the compensator, under an enlarged filtration, for the underlying source of risk in a return model. References include, among others, Karatzas and Pikovsky (1996), Grorud and Pontier (1998), Amendinger et al. (1998, 2003), Rindisbacher (1999), Ankirchner and Imkeller (2005) and Ankirchner et al. (2006).

This compensator plays a fundamental role in the theory of enlargements of filtrations (see Jeulin (1980), Jacod (1985), and Imkeller (1996) for an introduction). The terminology "(modified) price of risk" is used by Biagini and Oksendal (2005) to describe the MPR augmented by the PIPR. Ankirchner and Imkeller (2005) and Ankirchner et al. (2006) introduce the terminology "insider drift" which corresponds to the PIPR scaled by the volatility.

Section 2 examines the PIPR in discrete time models. It defines the notion and describes some of its prominent properties, provides illustrative examples and discusses its impact on the optimal portfolio of an informed investor. Section 3 focuses on continuous time models driven by Brownian uncertainty. It revisits the topics examined in the discrete time setting. It also provides some insights about tests for skill. Conclusions are in Section 4. Proofs of the main propositions can be found in the appendix.

2 The PIPR in discrete time

2.1 Definition and properties

Consider a setting with discrete dates $n \in \mathbb{N}$, a flow of public information $\mathbb{F} \equiv \{\mathcal{F}_n : n \in \mathbb{N}\}$, a risk free asset paying the predictable rate of return $r_n \in \mathcal{F}_n$ at $n+1$ and d risky assets paying the (vector of) risky rates of return $R_{n+1} \in \mathcal{F}_{n+1}$ at $n+1$. Assume that returns are square integrable (finite second moments) and that the covariance matrix $\Sigma_n^{\mathbb{F}} \equiv VAR\left[R_{n+1} | \mathcal{F}_n\right]$ has full rank. Under these conditions, excess returns $R_{n+1}^e \equiv R_{n+1} - r_n 1_d$ have the Doob-Meyer decomposition,

$$R_{n+1}^e = \sigma_n^{\mathbb{F}} \left(\theta_n^{\mathbb{F}} + \Delta W_n^{\mathbb{F}} \right) \tag{8.1}$$

relative to the public flow of information $\mathbb{F}$, where $\theta_n^{\mathbb{F}}$ is the market price of risk (MPR), $\sigma_n^{\mathbb{F}}$ is the return volatility and $\Delta W_n^{\mathbb{F}}$ is the d-dimensional vector of return innovations. These quantities are respectively given by,

$$\theta_n^{\mathbb{F}} = \left(\sigma_n^{\mathbb{F}}\right)^{-1} \left(E\left[R_{n+1} | \mathcal{F}_n\right] - r_n 1_d\right) \tag{8.2}$$

$$\sigma_n^{\mathbb{F}} \left(\sigma_n^{\mathbb{F}}\right)' = VAR\left[R_{n+1} | \mathcal{F}_n\right] \tag{8.3}$$

$$\Delta W_n^{\mathbb{F}} = \left(\sigma_n^{\mathbb{F}}\right)^{-1} \left(R_{n+1} - E\left[R_{n+1} | \mathcal{F}_n\right]\right). \tag{8.4}$$

The processes $\sigma^{\mathbb{F}}, \theta^{\mathbb{F}}$ are $\mathbb{F}$-adapted.[2] By construction, return innovations $\Delta W_n^{\mathbb{F}}$ are orthogonal $\mathbb{F}$-martingale difference sequences, $E\left[\Delta W_n^{\mathbb{F}} | \mathcal{F}_n\right] = 0$, with identity covariance matrix, $VAR\left[\Delta W_n^{\mathbb{F}} | \mathcal{F}_n\right] = I_d$. The MPR is the risk premium per unit risk in the public information flow.

Let us next examine the effects of private information on the structure of returns. To this end, consider an (informed) investor who receives a finer flow of information $\mathbb{G} \equiv \{\mathcal{G}_n : n \in \mathbb{N}\}$, where $\mathcal{G}_n \supset \mathcal{F}_n$ for each $n \in \mathbb{N}$. The flow $\mathbb{G}$ represents private information. The investor with private information $\mathbb{G}$ has better information than the market as he/she can distinguish more events and therefore pursue policies that are contingent on a refined set of states. Optimal decisions of the informed are adapted to $\mathbb{G}$, but not necessarily to $\mathbb{F}$ as he/she may act on events unknown to the public. The informed investor will therefore always be at least as well off as an otherwise identical agent with public information flow $\mathbb{F}$.

In order to find the representation of returns in the private information flow $\mathbb{G}$, note that the $\mathbb{F}$-return innovation can be decomposed as,

$$\Delta W_n^{\mathbb{F}} = \Delta W_n^{\mathbb{F}} - E\left[\Delta W_n^{\mathbb{F}} \middle| \mathcal{G}_n\right] + E\left[\Delta W_n^{\mathbb{F}} \middle| \mathcal{G}_n\right] \equiv \Delta W_n^{\mathbb{G},\mathbb{F}} + E\left[\Delta W_n^{\mathbb{F}} \middle| \mathcal{G}_n\right], \tag{8.5}$$

where the increment $\Delta W_n^{\mathbb{G},\mathbb{F}} \equiv \Delta W_n^{\mathbb{F}} - E\left[\Delta W_n^{\mathbb{F}} | \mathcal{G}_n\right]$ is a $\mathbb{G}$-innovation. The excess return can then be written as,

$$\begin{aligned} R_{n+1}^e &= E\left[R_{n+1}^e | \mathcal{G}_n\right] + R_{n+1}^e - E\left[R_{n+1}^e | \mathcal{G}_n\right] \\ &= \sigma_n^{\mathbb{F}} \left(\theta_n^{\mathbb{F}} + E\left[\Delta W_n^{\mathbb{F}} \middle| \mathcal{G}_n\right] + \Delta W_n^{\mathbb{G},\mathbb{F}}\right). \end{aligned} \tag{8.6}$$

By construction, the increment process $\left\{\Delta W_n^{\mathbb{G},\mathbb{F}} : n \in \mathbb{N}\right\}$ is an orthogonal, $\mathbb{G}$-martingale difference sequence, i.e., $E\left[\Delta W_n^{\mathbb{G},\mathbb{F}} \middle| \mathcal{G}_n\right] = E\left[\Delta W_n^{\mathbb{F}} | \mathcal{G}_n\right] - E\left[\Delta W_n^{\mathbb{F}} | \mathcal{G}_n\right] = 0$ for all $n \in \mathbb{N}$, with covariance matrix $VAR\left[\Delta W_n^{\mathbb{G},\mathbb{F}} \middle| \mathcal{G}_n\right] = \left(\sigma_n^{\mathbb{F}}\right)^{-1} \Sigma_n^{\mathbb{G}} \left(\left(\sigma_n^{\mathbb{F}}\right)'\right)^{-1}$ where $\Sigma_n^{\mathbb{G}} \equiv \sigma_n^{\mathbb{G}} \left(\sigma_n^{\mathbb{G}}\right)' \equiv VAR\left[R_{n+1} | \mathcal{G}_n\right]$ for $n \in \mathbb{N}$.

In the private information flow $\mathbb{G}$, the Doob-Meyer decomposition of asset excess returns is,

$$R_{n+1}^e = \sigma_n^{\mathbb{G}} \left(\theta_n^{\mathbb{G}} + \Delta W_n^{\mathbb{G}}\right) \tag{8.7}$$

for $\mathbb{G}$-adapted processes $\sigma^{\mathbb{G}}, \theta^{\mathbb{G}}$ and return innovation $\Delta W^{\mathbb{G}}$. Equation (8.7) expresses all return components with respect to the information flow $\mathbb{G}$. Thus,

$$\theta_n^{\mathbb{G}} = \left(\sigma_n^{\mathbb{G}}\right)^{-1} \left(E\left[R_{n+1} | \mathcal{G}_n\right] - r_n 1_d\right) \tag{8.8}$$

$$\sigma_n^{\mathbb{G}} \left(\sigma_n^{\mathbb{G}}\right)' = VAR\left[R_{n+1} | \mathcal{G}_n\right] \tag{8.9}$$

$$\Delta W_n^{\mathbb{G}} = \left(\sigma_n^{\mathbb{G}}\right)^{-1} \left(R_{n+1} - E\left[R_{n+1} | \mathcal{G}_n\right]\right). \tag{8.10}$$

It is of particular interest to note that the $\mathbb{G}$-volatility, $\sigma^{\mathbb{G}}$, differs from the $\mathbb{F}$-volatility, $\sigma^{\mathbb{F}}$.

Definition 1 (Private Information Price of Risk (PIPR)) *Let $\mathbb{F}, \mathbb{G}$ be ordered flows of information such that $\mathbb{F} \subset \mathbb{G}$ and let $R_{n+1}^e \equiv R_{n+1} - r_n 1_d$ be the vector of square integrable excess returns satisfying (8.1)–(8.4). The Private Information Price of Risk (PIPR) of the information flow $\mathbb{G}$ relative to the information flow $\mathbb{F}$, denoted by $\theta^{\mathbb{G},\mathbb{F}} \equiv \left\{\theta_n^{\mathbb{G},\mathbb{F}} : n \in \mathbb{N}\right\}$, is the predictable component of the Doob-Meyer decomposition of the $\mathbb{F}$-innovation $\Delta W^{\mathbb{F}}$ in the finer flow $\mathbb{G}$. That is,*

$$\theta_n^{\mathbb{G},\mathbb{F}} \equiv E\left[\Delta W_n^{\mathbb{F}} \middle| \mathcal{G}_n\right] \tag{8.11}$$

for $n \in \mathbb{N}$. Excess returns have the decomposition $R_{n+1}^e = \sigma_n^{\mathbb{F}} \left(\theta_n^{\mathbb{F}} + \theta_n^{\mathbb{G},\mathbb{F}} + \Delta W_n^{\mathbb{G},\mathbb{F}}\right)$ where $\Delta W_n^{\mathbb{G},\mathbb{F}} \equiv \Delta W_n^{\mathbb{F}} - \theta_n^{\mathbb{G},\mathbb{F}}$.

The PIPR $\theta^{\mathbb{G},\mathbb{F}}$ associated with the information flows $\mathbb{G},\mathbb{F}$, such that $\mathbb{G} \supseteq \mathbb{F}$, corresponds to the compensator of the $\mathbb{F}$-return innovation $\Delta W^{\mathbb{F}}$ in the private information $\mathbb{G}$. It represents the conditional mean of the $\mathbb{F}$-return innovation given $\mathbb{G}$: $\theta_n^{\mathbb{G},\mathbb{F}} \equiv E\left[\Delta W_n^{\mathbb{F}} | \mathcal{G}_n\right]$ for all $n \in \mathbb{N}$.

Given the definition of the PIPR and the decomposition (8.5), it follows that the $\mathbb{F}$-innovation in $\mathbb{G}$ satisfies $\Delta W_n^{\mathbb{G},\mathbb{F}} = \Delta W_n^{\mathbb{F}} - \theta_n^{\mathbb{G},\mathbb{F}}$ and that excess returns can be written as,

$$R_{n+1}^e = \sigma_n^{\mathbb{F}} \left(\theta_n^{\mathbb{F}} + \theta_n^{\mathbb{G},\mathbb{F}} + \Delta W_n^{\mathbb{G},\mathbb{F}}\right). \tag{8.12}$$

This expression shows that the PIPR can be interpreted as the incremental price of risk, relative to $\theta_n^{\mathbb{F}}$, due to the availability of the private information $\mathbb{G}$. This incremental price of risk is measured for the total risk, $\sigma_n^{\mathbb{F}}$, calculated under information $\mathbb{F}$.

The connection with the return structure (8.7) in the $\mathbb{G}$-information is also instructive. Simple algebra shows that,

$$\theta_n^{\mathbb{G}} = \left(\sigma_n^{\mathbb{G}}\right)^{-1} E\left[R_{n+1}^e | \mathcal{G}_n\right] = \left(\sigma_n^{\mathbb{G}}\right)^{-1} \sigma_n^{\mathbb{F}} \left(\theta_n^{\mathbb{F}} + \theta_n^{\mathbb{G},\mathbb{F}}\right)$$

$$\Delta W_n^{\mathbb{G}} = \left(\sigma_n^{\mathbb{G}}\right)^{-1} \left(R_{n+1}^e - E\left[R_{n+1}^e | \mathcal{G}_n\right]\right) = \left(\sigma_n^{\mathbb{G}}\right)^{-1} \sigma_n^{\mathbb{F}} \Delta W_n^{\mathbb{G},\mathbb{F}}.$$

Hence, the $\mathbb{G}$-price of risk $\theta^{\mathbb{G}}$ is a normalized version of the total price of risk $\theta_n^{\mathbb{F}} + \theta_n^{\mathbb{G},\mathbb{F}}$ for the informed associated with the risk exposure $\sigma^{\mathbb{F}}$. Likewise, the $\mathbb{G}$-return innovation $\Delta W_n^{\mathbb{G}}$ is a normalized version of the $\mathbb{G}$-innovation $\Delta W_n^{\mathbb{G},\mathbb{F}} \equiv \Delta W_n^{\mathbb{F}} - \theta_n^{\mathbb{G},\mathbb{F}}$. If risk exposures coincide, i.e., $\sigma_n^{\mathbb{G}} = \sigma_n^{\mathbb{F}}$, then $\theta_n^{\mathbb{G}} = \theta_n^{\mathbb{F}} + \theta_n^{\mathbb{G},\mathbb{F}}$ and $\Delta W_n^{\mathbb{G}} = \Delta W_n^{\mathbb{G},\mathbb{F}}$.

The following properties follow from the definition and relations above.

Proposition 2 (Properties of PIPR:) *The PIPR has the following properties:*

(i) ***No-arbitrage:*** $E\left[\theta_n^{\mathbb{G},\mathbb{F}} \middle| \mathcal{F}_n\right] = 0$

(ii) ***Incremental price of risk:*** $\theta_n^{\mathbb{G},\mathbb{F}} = \left(\sigma_n^{\mathbb{F}}\right)^{-1} \sigma_n^{\mathbb{G}} \theta_n^{\mathbb{G}} - \theta_n^{\mathbb{F}}$. *For return innovations that are conditionally homoskedastic* $\sigma_n^{\mathbb{G}} = \sigma_n^{\mathbb{F}} = \sigma$ *for some constant* $d \times d$ *matrix* σ, $\theta_n^{\mathbb{G},\mathbb{F}} = \theta_n^{\mathbb{G}} - \theta_n^{\mathbb{F}}$, *i.e., the PIPR corresponds to the difference in prices of risk relative to the flows of information* $\mathbb{G}$ *and* $\mathbb{F}$.

(iii) ***Covariance with density process:*** *If* $\mathcal{G}_n = \mathcal{F}_n \bigvee \sigma\,(G)$ *for some m-dimensional random vector* $G \in \mathcal{F}_N$,[3]

$$\theta_n^{\mathbb{G},\mathbb{F}} = E\left[\Delta W_n^{\mathbb{F}} \frac{P(G \in dx | \mathcal{F}_{n+1})}{P(G \in dx | \mathcal{F}_n)} \middle| \mathcal{F}_n\right]_{|x=G} \equiv \theta_n^{G|F}(x)_{|x=G}. \qquad (8.13)$$

(iv) ***Representation of density process of conditional signal:*** *The conditional density process* $Z_{n,n+1}^{G}(x) \equiv P(G \in dx | \mathcal{F}_{n+1}) / P(G \in dx | \mathcal{F}_n)$ *has the representation*

$$Z_{n,n+1}^{G}(x) = 1 + \theta_n^{G|F}(x)' \Delta W_n^{\mathbb{F}} + \Delta V_n^{\mathbb{F}}(x) \qquad (8.14)$$

where $\Delta V^{\mathbb{F}}(x)$ *is an* $\mathbb{F}$*-martingale difference sequence orthogonal to* $\Delta W^{\mathbb{F}}$ *for any* $x \in \mathbb{R}^m$, *i.e.,* $E\left[\Delta V_n^{\mathbb{F}}(x) | \mathcal{F}_n\right] = E\left[\Delta V_n^{\mathbb{F}}(x)\, \Delta W_n^{\mathbb{F}} | \mathcal{F}_n\right] = 0$ *for all* $n \in \mathbb{N}$.

The first property is a no-arbitrage condition. The intuition is straightforward. By the tower property of conditional expectations $E\left[\Delta W_n^{\mathbb{G},\mathbb{F}} \middle| \mathcal{F}_n\right] = E\left[E\left[\Delta W_n^{\mathbb{G},\mathbb{F}} \middle| \mathcal{G}_n\right] \middle| \mathcal{F}_n\right] = 0$. Therefore,

$$E\left[R_{n+1}^e | \mathcal{F}_n\right] = E\left[\sigma_n^{\mathbb{F}}\left(\theta_n^{\mathbb{F}} + \theta_n^{\mathbb{G},\mathbb{F}} + \Delta W_n^{\mathbb{G},\mathbb{F}}\right) \middle| \mathcal{F}_n\right] = \sigma_n^{\mathbb{F}}\left(\theta_n^{\mathbb{F}} + E\left[\theta_n^{\mathbb{G},\mathbb{F}} \middle| \mathcal{F}_n\right]\right).$$

To ensure the absence of arbitrage opportunities under $\mathbb{F}$, the $\mathbb{F}$-expectation of the excess return must be equal to $\sigma_n^{\mathbb{F}} \theta_n^{\mathbb{F}}$. Thus, $E\left[\theta_n^{\mathbb{G},\mathbb{F}} \middle| \mathcal{F}_n\right] = 0$. In other words, the uninformed (i.e., publicly informed) investor cannot perceive an expected excess return that deviates from $\theta_n^{\mathbb{F}}$. A consequence of this property is that the PIPR is a martingale with null expectation in the public information flow $\mathbb{F}$. Moreover, the standard measure Q is an equivalent martingale measure and $E^Q\left[\theta_n^{\mathbb{F}} + \Delta W_n^{\mathbb{F}} | \mathcal{F}_n\right] = E\left[\Delta W_n^{\mathbb{F}} | \mathcal{F}_n\right] = 0$. The risk neutral formula $E^Q\left[R_{n+1} | \mathcal{F}_n\right] = r_n$ follows.

The second property is also discussed before the proposition. The volatility matrix $\sigma_n^{\mathbb{F}}$ measures the risk exposure of the excess return given public information. The expression $\sigma_n^{\mathbb{F}}\theta^{\mathbb{G},\mathbb{F}} = \sigma_n^{\mathbb{G}}\theta_n^{\mathbb{G}} - \sigma_n^{\mathbb{F}}\theta_n^{\mathbb{F}}$ represents the incremental risk premium due to private information. The PIPR therefore measures the incremental price of risk, due to private information, for $\mathbb{F}$-risk exposure $\sigma_n^{\mathbb{F}}$.

The third property is a consequence of Bayes' law. By definition $\theta_n^{\mathbb{G},\mathbb{F}} = E\left[\Delta W_n^{\mathbb{F}} \middle| \mathcal{G}_n\right]$, where the expectation is calculated relative to the conditional probability given $G, \mathcal{F}_n$, $P\left(\Delta W_n^{\mathbb{F}} \in dw \middle| G = x, \mathcal{F}_n\right)$. Bayes' law shows that,

$$P\left(\Delta W_n^{\mathbb{F}} \in dw \middle| G = x, \mathcal{F}_n\right) = \frac{E\left[P(G \in dx|\mathcal{F}_{n+1})|\,\Delta W_n^{\mathbb{F}} = w, \mathcal{F}_n\right]}{P(G \in dx|\mathcal{F}_n)} P\left(\Delta W_n^{\mathbb{F}} \in dw \middle| \mathcal{F}_n\right)$$

(see appendix for details) implying,

$$\begin{aligned} E\left[\Delta W_n^{\mathbb{F}} \middle| \mathcal{G}_n\right] &= E\left[\Delta W_n^{\mathbb{F}} \frac{E\left[P(G \in dx|\mathcal{F}_{n+1})|\,\Delta W_n^{\mathbb{F}} = w, \mathcal{F}_n\right]}{P(G \in dx|\mathcal{F}_n)} \middle| \mathcal{F}_n\right] \\ &= E\left[\Delta W_n^{\mathbb{F}} \frac{P(G \in dx|\mathcal{F}_{n+1})}{P(G \in dx|\mathcal{F}_n)} \middle| \mathcal{F}_n\right] \\ &= COV\left[\Delta W_n^{\mathbb{F}}, \frac{\Delta P(G \in dx|\mathcal{F}_n)}{P(G \in dx|\mathcal{F}_n)} \middle| \mathcal{F}_n\right] \end{aligned}$$

where $\Delta P(G \in dx|\mathcal{F}_n) = P(G \in dx|\mathcal{F}_{n+1}) - P(G \in dx|\mathcal{F}_n)$. In essence, the $\mathbb{G}$-expectation is an $\mathbb{F}$-expectation relative to an adjusted probability measure. More precise information causes the investor to adjust the likelihood of the various events by the ratio,

$$Z_{n,n+1}^{G}(x) = \frac{P(G \in dx|\mathcal{F}_{n+1})}{P(G \in dx|\mathcal{F}_n)}.$$

Property (iii) suggests a straightforward procedure to calculate the PIPR. If the conditional density process of the signal $\left\{Z_{n,n+1}^{G}(x) : n \in \mathbb{N}\right\}$ is known, the PIPR can be calculated as the conditional covariance between the density process and the return innovation under the public information $\mathbb{F}$.[4] In particular, if the signal G becomes public information at $n+1$, i.e., $G \in \mathcal{F}_{n+1}$, then,

$$\theta_n^{G|F}(x) = E\left[\Delta W_n^{\mathbb{F}} Z_{n,n+1}^{G}(x) \middle| \mathcal{F}_n\right] = E\left[\Delta W_n^{\mathbb{F}} \middle| \mathcal{F}_n\right] \frac{\delta_x(G)}{P(G \in dx|\mathcal{F}_n)} dx = 0$$

where $\delta_x(G)$ is the Dirac delta with mass at x. Public knowledge of the signal at $n+1$ implies that it has no value at n: the PIPR is null.

Property (iv) shows that the PIPR $\theta_n^{\mathbb{G},\mathbb{F}} = \theta_n^{G|F}(G)$ is the integrand of the martingale representation of the conditional density process in terms of the return innovation $\Delta W_n^{\mathbb{F}}$ and the orthogonal martingale difference sequence $\Delta V_n^{\mathbb{F}}$. If markets are complete, the only martingale difference sequence orthogonal to the return innovation is $\Delta V_n^{\mathbb{F}} = 0$. In this case, the conditional density process of the signal has the representation $Z_{n,n+1}^{G}(x) = 1 + \theta_n^{G|F}(x)' \Delta W_n^{\mathbb{F}}$. In discrete time, the model has complete markets if and only if the information algebra $\mathcal{F}_n$ has at most $(d+1)^n$ atoms.[5] In continuous time, the market is complete if and only if $\sigma_n^{\mathbb{F}}$ has full rank. In this case, as will be shown in Section 3, there exists a multiplicative martingale representation of the conditional density process of the form $dZ_{t,s}^{G} = Z_{t,s}^{G}\theta_s^{G|F}(x)\,dW_s^{\mathbb{F}}$ and the PIPR is the covariation between the log-density process and the return innovation.

2.2 Examples

Example 1: Binomial model

Assume that $r_n = r$, a constant. Consider a recombining binomial tree over two periods (3 dates) with state space $\Omega = \{u,d\}^2$. Public information flow is $\mathcal{F}_0 = \{\Omega,\emptyset\}$, $\mathcal{F}_1 = \left\{\Omega,\emptyset,\left\{u^2,ud\right\},\left\{d^2,du\right\}\right\}$ and $\mathcal{F}_2 = 2^{\Omega}$, the set of all subsets of Ω. The filtration is generated by the stock price, which follows a random walk driven by return innovations $S_n = S_0\prod_{j=0}^{n-1} X_{j+1}$ where X_j is an i.i.d sequence of binary variables with probability density $P\left(X_{j+1}=u\middle|\mathcal{F}_j\right)=p$, $P\left(X_{j+1}=d\middle|\mathcal{F}_j\right) = 1-p$. The stock's rates of return are $R_{n+1} = S_{n+1}/S_n - 1 = X_{n+1} - 1$ and therefore i.i.d.. As $E\left[R_{n+1}|\mathcal{F}_n\right] = (up+d(1-p))-1$ and $VAR\left[R_{n+1}|\mathcal{F}_n\right] = VAR\left[X_{n+1}|\mathcal{F}_n\right] = p(1-p)(u-d)^2$, excess returns have the representation (8.1) with,

$$\sigma_n = \frac{1}{2}\sqrt{p(1-p)}\,(u-d), \qquad \theta_n^{\mathbb{F}} = \frac{pu+(1-p)d-(1+r)}{\sqrt{p(1-p)}\,(u-d)}$$

$$\Delta W_n^{\mathbb{F}} = \frac{X_{n+1}-(pu+(1-p)d)}{\sqrt{p(1-p)}(u-d))}.$$

The volatility process is constant, $\sigma_n = \sigma$. The market price of risk is also constant, $\theta_n^{\mathbb{F}} = \theta^{\mathbb{F}}$.

Suppose that the private signal is $G = S_2$, thus reveals the stock price at date 2. It follows that,

$$\begin{aligned}\theta_0^{\mathbb{G},\mathbb{F}} = E\left[\Delta W_0^{\mathbb{F}}\middle|\mathcal{G}_0\right] &= \frac{E\left[X_1|\{X_1=u,X_2=d\}\cup\{X_1=d,X_2=u\}\right]}{\sqrt{p(1-p)}(u-d)}1_{S_0ud}(G)\\ &+\frac{u}{\sqrt{p(1-p)}(u-d)}1_{S_0u^2}(G)+\frac{d}{\sqrt{p(1-p)}(u-d)}1_{S_0d^2}(G)-\frac{pu+(1-p)d}{\sqrt{p(1-p)}(u-d)}\\ &=\frac{\frac{1}{2}(u+d)1_{S_0ud}(G)+u1_{S_0u^2}(G)+d1_{S_0d^2}(G)-pu-(1-p)d}{\sqrt{p(1-p)}(u-d)}\end{aligned}$$

and therefore,

$$\theta_0^{\mathbb{G},\mathbb{F}} = \frac{\left(\frac{1}{2}-p\right)u+\left(\frac{1}{2}-(1-p)\right)d}{\sqrt{p(1-p)}\,(u-d)}1_{S_0ud}(G)+\sqrt{\frac{1-p}{p}}1_{S_0u^2}(G)-\sqrt{\frac{p}{1-p}}1_{S_0d^2}(G). \tag{8.15}$$

For interpretation purposes, it is useful to distinguish two cases.

If $p=1/2$ then $\theta_0^{\mathbb{G},\mathbb{F}} = 1_{S_0u^2}(G)-1_{S_0d^2}(G)$. Hence, the PIPR at time zero is null if $G=S_2=S_0ud$, one if $G=S_2=S_0u^2$ and minus one if $G=S_2=S_0d^2$. As the return innovations are symmetric, the PIPR differs from zero only if the realization of the signal is informative about the realization of the initial return innovation. This is the case if the stock price goes up in both periods and hence the signal reveals that $X_1 = u$ implying the realization $\Delta W_0^{\mathbb{F}} = \sqrt{\frac{1-p}{p}} = 1$ of the return innovation. If the stock price goes down twice, the signal reveals that $X_1 = d$ implying the realization $\Delta W_0^{\mathbb{F}} = -\sqrt{\frac{p}{1-p}} = -1$ of the return innovation.

In all other cases, i.e., if $G = S_2 = S_0ud$, the realization of the signal does not reveal anything about the realization of X_1 and therefore about the return innovation $\Delta W_0^{\mathbb{F}}$. It follows that the PIPR is null.

If $p \neq 1/2$, the increments X_n do not have a symmetric distribution. In this case, the PIPR is positive if $G = S_0u^2$ revealing that $X_1 = u$ and negative if $G = S_0d^2$ revealing that $X_1 = d$. If $G = S_0ud$, the PIPR is positive if $p > 1/2$ and negative if $p < 1/2$. If $p \ll 1/2$, the state $S_2 = S_0u^2$ is less likely. Therefore, knowing that $G = S_0u^2$, so that $X_1 = u$, is more valuable and results in a large positive PIPR. Similarly, if $p \gg 1/2$ the likelihood of the down state is low. Knowing that $G = S_0d^2$, hence that $X_1 = d$, is therefore more valuable and induces a large negative PIPR.

Finally, it is interesting to note that the realization of the PIPR does not depend on the magnitude of the price change if the signal fully reveals the realization of X_1, i.e., if $G = S_0u^2$ or $G = S_0d^2$.

The PIPR at time one is,

$$\begin{aligned}
\theta_1^{\mathbb{G},\mathbb{F}} &= E\left[\Delta W_1^{\mathbb{F}} \middle| \mathcal{G}_1\right] = E\left[\Delta W_1^{\mathbb{F}} \middle| S_1 = y, y\left(1 + r + \sigma\left(\theta^{\mathbb{F}} + \Delta W_1^{\mathbb{F}}\right)\right) = x\right]_{|x=G,y=S_1} \\
&= E\left[\Delta W_1^{\mathbb{F}} \middle| S_1 = y, \Delta W_1^{\mathbb{F}} = \frac{\frac{x}{y} - (1+r)}{\sigma} - \theta^{\mathbb{F}}\right]_{|x=G,y=S_1} = \frac{\frac{G}{S_1} - (1+r)}{\sigma} - \theta^{\mathbb{F}} \\
&= \frac{R_2 - E[R_2|\mathcal{F}_1]}{\sqrt{VAR[R_2|\mathcal{F}_1]}} = \Delta W_1^{\mathbb{F}}.
\end{aligned}$$

The PIPR at time one corresponds to the innovation relative to public information.[6] The sign of the PIPR is the sign of the innovation. As $\theta_1^{\mathbb{G},\mathbb{F}} = \Delta W_1^{\mathbb{F}}$, the Doob-Meyer decomposition of the return during the last period becomes,

$$R_2^e = \sigma\left(\theta^{\mathbb{F}} + \Delta W_1^{\mathbb{F}}\right) = \sigma\left(\theta^{\mathbb{F}} + \theta_1^{\mathbb{G},\mathbb{F}}\right)$$

and the innovation for private information vanishes, $\Delta W_1^{\mathbb{G},\mathbb{F}} = \Delta W_1^{\mathbb{F}} - \theta_1^{\mathbb{G},\mathbb{F}} = 0$. This follows, because knowledge of the signal $G = S_2$ and of the price S_1 renders the return $R_2 = S_2/S_1 - 1$ predictable, i.e. $R_2 \in \mathcal{G}_1$. There is no surprise in the risky asset price change.

Example 2: Gaussian innovation

Consider a setting with conditionally independent Gaussian innovations, $P\left(\Delta W_n^{\mathbb{F}} \in dx \middle| \mathcal{F}_n\right) = \phi(x)\,dx$ where $\phi(x) = \exp\left(-x^2/2\right)/\sqrt{2\pi}$ denotes the Gaussian density. Assume that $r_n = r$ constant and that the stock price process is $S_{n+1} =$

$S_n \exp(R_{n+1})$ where the continuously compounded return is $\log(S_{n+1}/S_n) = R_{n+1} = r + \sigma\left(\theta^{\mathbb{F}} + \Delta W_n^{\mathbb{F}}\right)$ with constant volatility σ and MPR $\theta^{\mathbb{F}}$. First assume that $G = h(S_N)$ for some measurable one-to-one mapping $h : \mathbb{R} \to \mathbb{R}$. In this case,

$$\begin{aligned} P(G \leq x | \mathcal{F}_n) &= P\left(\sigma \sum_{j=n}^{N-1} W_{j+1}^{\mathbb{F}} \leq \log\left(h^{-1}(x)/S_n\right) - \left(r + \sigma\theta^{\mathbb{F}}\right)(N-n) \middle| \mathcal{F}_n \right) \\ &= \int_{-\infty}^{d\left(\frac{h^{-1}(x)}{S_n}, N-n\right)} \phi(u)\, du = \Phi\left(d\left(\frac{h^{-1}(x)}{S_n}, N-n \right) \right) \end{aligned}$$

where,

$$d\left(\frac{h^{-1}(x)}{S_n}, N-n \right) = \frac{\log\left(h^{-1}(x)/S_n\right) - \left(r + \sigma\theta^{\mathbb{F}}\right)(N-n)}{\sigma\sqrt{N-n}} \equiv \frac{\alpha(x, S_n, N-n)}{\sigma\sqrt{N-n}}$$

and $\Phi(\cdot)$ is the standard normal cumulative distribution function. The density function is,

$$P(G \in dx | \mathcal{F}_n) = \phi\left(d\left(\frac{h^{-1}(x)}{S_n}, N-n \right) \right) \frac{1}{\sigma\sqrt{N-n}} \left| \frac{\partial_x h^{-1}(x)}{h^{-1}(x)} \right|.$$

Applying (iii) of Proposition 2 and using,

$$\begin{aligned} \alpha(x, S_{n+1}, N-(n+1)) &= \log h^{-1}(x) - \log S_{n+1} - \left(r + \sigma\theta^{\mathbb{F}}\right)(N-(n+1)) \\ &= \log h^{-1}(x) - \log S_n - \left(r + \sigma\theta^{\mathbb{F}}\right)(N-n) - \sigma\,\Delta W_n^{\mathbb{F}} \\ &= \alpha(x, S_n, N-n) - \sigma\,\Delta W_n^{\mathbb{F}} \\ \frac{P(G \in dx | \mathcal{F}_{n+1})}{P(G \in dx | \mathcal{F}_n)} &= \frac{\sqrt{N-n}}{\sqrt{N-(n+1)}} \frac{\phi\left(\frac{\alpha(x, S_{n+1}, N-(n+1))}{\sigma\sqrt{N-(n+1)}} \right)}{\phi\left(\frac{\alpha(x, S_n, N-n)}{\sigma\sqrt{N-n}} \right)} \\ &= \frac{\sqrt{N-n}}{\sqrt{N-(n+1)}} \frac{\phi\left(\frac{\alpha(x, S_n, N-n) - \sigma\,\Delta W_n^{\mathbb{F}}}{\sigma\sqrt{N-(n+1)}} \right)}{\phi\left(\frac{\alpha(x, S_n, N-n)}{\sigma\sqrt{N-n}} \right)} \\ \left(\frac{w - \alpha(x, S_n, N-n)/\sigma}{\sqrt{N-(n+1)}} \right)^2 + w^2 &= \frac{N-n}{N-(n+1)} \left(w - \frac{\alpha(x, S_n, N-n)}{\sigma(N-n)} \right)^2 \\ &\quad + \left(\frac{\alpha(x, S_n, N-n)}{\sigma\sqrt{N-n}} \right)^2 \end{aligned}$$

gives,

$$\begin{aligned}
\theta_n^{G|F}(x) &= \frac{\sqrt{N-n}}{\sqrt{N-(n+1)}} E\left[\Delta W_n^{\mathbb{F}} \frac{\phi\left(\frac{\alpha(x,S_n,N-n)-\sigma\,\Delta W_n^{\mathbb{F}}}{\sigma\sqrt{N-(n+1)}}\right)}{\phi\left(\frac{\alpha(x,S_n,N-n)}{\sigma\sqrt{N-n}}\right)} \middle| \mathcal{F}_n \right] \\
&= \frac{\sqrt{N-n}}{\sqrt{N-(n+1)}} \frac{1}{\phi\left(\frac{\alpha(x,S_n,N-n)}{\sigma\sqrt{N-n}}\right)} E\left[\Delta W_n^{\mathbb{F}} \phi\left(\frac{\alpha\,(x,S_n,N-n)-\sigma\,\Delta W_n^{\mathbb{F}}}{\sigma\sqrt{N-(n+1)}}\right) \middle| \mathcal{F}_n \right] \\
&= \frac{\sqrt{N-n}}{\sqrt{N-(n+1)}} \frac{1}{2\pi\,\phi\left(\frac{\alpha(x,S_n,N-n)}{\sigma\sqrt{N-n}}\right)} \int_{-\infty}^{\infty} w e^{-\frac{1}{2}\left(\frac{w-\alpha(x,S_n,N-n)/\sigma}{\sqrt{N-(n+1)}}\right)^2-\frac{1}{2}w^2} dw \\
&= \frac{\sqrt{N-n}}{\sqrt{N-(n+1)}} \frac{1}{2\pi\,\phi\left(\frac{\alpha(x,S_n,N-n)}{\sigma\sqrt{N-n}}\right)} \int_{-\infty}^{\infty} \\
&\quad w e^{-\frac{1}{2}\frac{N-n}{N-(n+1)}\left(w-\frac{\alpha(x,S_n,N-n)}{\sigma(N-n)}\right)^2-\frac{1}{2}\left(\frac{\alpha(x,S_n,N-n)}{\sigma\sqrt{N-n}}\right)^2} dw \\
&= \frac{\sqrt{N-n}}{\sqrt{N-(n+1)}} \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} w e^{-\frac{1}{2}\frac{N-n}{N-(n+1)}\left(w-\frac{\alpha(x,S_n,N-n)}{\sigma(N-n)}\right)^2} dw = \frac{\alpha\,(x,S_n,N-n)}{\sigma\,(N-n)}.
\end{aligned}$$

The PIPR, for $G = h(S_N)$, is,

$$\theta_n^{\mathbb{G},\mathbb{F}} = \frac{\alpha\,(S_N,S_n,N-n)}{\sigma\,(N-n)} = \frac{\log(S_N/S_n) - \left(r+\sigma\theta^{\mathbb{F}}\right)(N-n)}{\sigma\,(N-n)}.$$

It is positive (negative) if the realization of the compounded rate of return is above (below) its expected value.

For $n = N-1$, the PIPR becomes,

$$\theta_n^{\mathbb{G},\mathbb{F}} = \frac{\alpha\,(S_N,S_{N-1},1)}{\sigma} = \frac{\log(S_N/S_{N-1}) - \left(r+\sigma\theta^{\mathbb{F}}\right)}{\sigma} = \Delta W_1^{\mathbb{F}}$$

and the innovation for private information vanishes, $\Delta W_1^{\mathbb{G},\mathbb{F}} = \Delta W_1^{\mathbb{F}} - \theta_1^{\mathbb{G},\mathbb{F}} = 0$. The private signal $G = S_N$ and public information S_{N-1} reveal the return $R_N = \log(S_N/S_{N-1})$, i.e., $R_{N-1} \in \mathcal{G}_{N-1}$. There is perfect foresight about the next return realization. Using Bayes' rule, the conditional probability of the return innovation given the realization of the signal becomes,

$$\begin{aligned}
P\left(\Delta W_{N-1}^{\mathbb{F}} \in dw \middle| G = x \right) &= \frac{P\left(G \in dx |\, \Delta W_{N-1}^{\mathbb{F}} = w\right)}{P(G \in dx)} P\left(\Delta W_{N-1}^{\mathbb{F}} \in dw\right) \\
&= \frac{\delta\left(\log x - S_{N-1} - \left(r+\sigma\theta^{\mathbb{F}}+w\right)\right)}{\phi\left(\frac{\alpha(x,S_0)}{\sigma\sqrt{N}}\right)\frac{|\partial_x\alpha(x,S_0)|}{\sigma\sqrt{N}}} \phi(w)\,dw
\end{aligned}$$

where $\delta(\cdot)$ is the Dirac-delta function at zero. This conditional probability measure is singular. Given the local singularity and the absence of risk, an equivalent martingale measure cannot exist on the enlarged flow of information $\mathbb{G}$. From the fundamental theorem of asset pricing, establishing the equivalence between the absence of arbitrage opportunities and the existence of an equivalent martingale, it then follows that an informed investor can find arbitrage opportunities. The optimal portfolio choice problem is ill-posed and does not have a solution.[7]

Singularities and arbitrage opportunities can be avoided by adding noise to the signal. If $G = h(S_N) + \zeta$ for some independent random variable ζ with density $f_\zeta(\cdot)$,

$$P\left(\Delta W_{N-1}^{\mathbb{F}} \in dw \middle| G = x \right) = \frac{f_\zeta\left(\log x - S_{N-1} - \left(r + \sigma\theta^{\mathbb{F}} + w\right)\right)}{\phi\left(\frac{\alpha(x,S_0)}{\sigma\sqrt{N}}\right)\frac{|\partial_x \alpha(x,S_0)|}{\sigma\sqrt{N}}} \phi(w)\,dw.$$

In this case, the conditional probability of the return innovation given the signal remains equivalent to the unconditional probability. An equivalent martingale measure exists and the portfolio choice problem of the informed is well defined. Information remains sufficiently imprecise to preclude arbitrage opportunities even in the last period.

2.3 Optimal informed portfolio and PIPR

In order to illustrate the relation between the optimal portfolio of an informed investor and the PIPR, we focus on an extension of the standard setting in informational economics (e.g., Grossman and Stiglitz (1980)) to arbitrary distributions. The model has one period (two dates) and $d+1$ assets, d risky stock and a riskless asset. The rate of return on the riskless asset is r. The vector of excess returns on the stocks is $R_1^e = \sigma_0^{\mathbb{F}}\left(\theta_0^{\mathbb{F}} + \Delta W_0^{\mathbb{F}}\right)$ where $\Delta W_0^{\mathbb{F}}$ is a d-dimensional random variable with arbitrary distribution and $\theta_0^{\mathbb{F}}$ is the d-dimensional market price of risk. The coefficient $\sigma_0^{\mathbb{F}}$ is a $d \times d$ matrix of risk exposures, which is assumed to be invertible. There are no restrictions on asset positions. Public information is the trivial algebra.

The informed investor has information set $\mathbb{G} = \mathcal{G}_0$ and von Neumann-Morgenstern preferences with constant absolute risk averse utility function. Thus, $U_0^{\mathbb{G}} = -E\left[\exp(-AX_1)|\mathcal{G}_0\right]$ where $A > 0$ is the absolute risk aversion coefficient and X_1 is terminal wealth. Initial wealth is X_0. The informed invests π_0 in the risky stocks and the remainder $X_0 - \pi_0' 1$ in the riskless asset. Thus, terminal wealth is $X_1 = X_0(1+r) + \pi_0'\sigma_0^{\mathbb{F}}\left(\theta_0^{\mathbb{F}} + \theta_0^{\mathbb{G},\mathbb{F}} + \Delta W_0^{\mathbb{G},\mathbb{F}}\right)$, where excess returns are expressed relative to the private information.

Substituting final wealth in the utility function gives,

$$U_0^{\mathbb{G}} = -\exp\left(-A\left(X_0(1+r) + \pi_0'\sigma_0^{\mathbb{F}}\left(\theta_0^{\mathbb{F}} + \theta_0^{\mathbb{G},\mathbb{F}}\right)\right) + \kappa_0\left(-A\left(\sigma_0^{\mathbb{F}}\right)'\pi_0\right)\right)$$

where $\kappa_0(\tau) \equiv \log E\left[\exp\left(\tau'\Delta W_0^{\mathbb{G},\mathbb{F}}\right)\middle|\mathcal{G}_0\right]$ denotes the cumulant generating function of the return innovation for the enlarged information set $\mathbb{G}$. Maximizing expected utility is therefore equivalent to solving,

$$\max_{\pi_0 \in \mathcal{G}_0} \left\{ A\left(\pi_0' \sigma_0^{\mathbb{F}} \left(\theta_0^{\mathbb{F}} + \theta_0^{\mathbb{G},\mathbb{F}}\right)\right) - \kappa_0 \left(-A\left(\sigma_0^{\mathbb{F}}\right)' \pi_0\right)\right\}.$$

The cumulant generating function is a strictly convex function. The optimal portfolio of the informed is therefore,

$$\pi_0^{\mathbb{G}} = \frac{1}{A}\left(\left(\sigma_0^{\mathbb{F}}\right)'\right)^{-1} J_0\left(-\left(\theta_0^{\mathbb{F}} + \theta_0^{\mathbb{G},\mathbb{F}}\right)\right) \tag{8.16}$$

where $J_0(x) = -[\partial\kappa_0]^{-1}(x)$ denotes the inverse of the derivative of the cumulant generating function.[8]

Suppose that the conditional distribution of the innovation $\Delta W_0^{\mathbb{G},\mathbb{F}}$ is given by an exponential family with natural parameterization ϑ. In this case, the gradient of the cumulant generating function is connected to the mean of the sufficient statistic for the parameters of the distribution, $\partial\kappa_0(\tau) = E_{\tau+\vartheta}\left[T\left(\Delta W_0^{\mathbb{G},\mathbb{F}}\right)\middle|\mathcal{G}_0\right]$, where the right hand side is the expected value of the sufficient statistic $T(\cdot)$ for the parameter ϑ, under the local parameter shift $\vartheta+\tau$. It follows that the optimal portfolio is obtained by evaluating the negative of the inverse of the parameter map $\tau \mapsto E_{\vartheta+\tau}\left[T\left(\Delta W_0^{\mathbb{G},\mathbb{F}}\right)\middle|\mathcal{G}_0\right]$ at $-\left(\theta_0^{\mathbb{F}} + \theta_0^{\mathbb{G},\mathbb{F}}\right)$. It is therefore fully determined by the parameters of the mean of the sufficient statistic for the parameter of the innovation.

For conditionally Gaussian innovation, as in the standard informational economics model, the cumulant generating function is quadratic, $\kappa_0(x) = \frac{1}{2}\|x\|^2$, and $J_0\left(-\left(\theta_0^{\mathbb{F}} + \theta_0^{\mathbb{G},\mathbb{F}}\right)\right) = \theta_0^{\mathbb{F}} + \theta_0^{\mathbb{G},\mathbb{F}}$ is linear. Alternatively, a sufficient statistic for the mean $\vartheta = 0$ is $T\left(\Delta W_0^{\mathbb{G},\mathbb{F}}\right) = \Delta W_0^{\mathbb{G},\mathbb{F}}$ and $E_{\vartheta+\tau}\left[T\left(\Delta W_0^{\mathbb{G},\mathbb{F}}\right)\right] = \tau$. Evaluating the negative of the inverse map at the point $-\left(\theta_0^{\mathbb{F}} + \theta_0^{\mathbb{G},\mathbb{F}}\right)$ also gives $\theta_0^{\mathbb{F}} + \theta_0^{\mathbb{G},\mathbb{F}}$. The resulting optimal portfolio of the informed is,

$$\pi_0^{\mathbb{G}} = \frac{1}{A}\left(\left(\sigma_0^{\mathbb{F}}\right)'\right)^{-1} \left(\theta_0^{\mathbb{F}} + \theta_0^{\mathbb{G},\mathbb{F}}\right) \equiv \pi_0^{\mathbb{F}} + \pi_0^{\mathbb{G},\mathbb{F}}. \tag{8.17}$$

It has a mean-variance structure, with two components. The component $\pi_0^{\mathbb{F}}$ represents the asset allocation based on public information. The component $\pi_0^{\mathbb{G},\mathbb{F}}$ is the incremental allocation due to private information. The latter is fully determined by the PIPR. It vanishes when private information has no value ($\theta_0^{\mathbb{G},\mathbb{F}} = 0$).

3 The PIPR in continuous time

3.1 Definition and properties

Consider a typical continuous time setting where uncertainty is described by a standard Brownian motion $W^{\mathbb{F}}$ defined on a complete probability space

$(\Omega, \mathcal{F}, P)$. The public filtration $\mathbb{F}$ is the filtration generated by $W^{\mathbb{F}}$. Consider a risky asset with price evolving according to,

$$\frac{dS_t}{S_t} = r_t dt + \sigma_t \left(\theta_t^{\mathbb{F}} dt + dW_t^{\mathbb{F}}\right), \quad S_0 \text{ given} \tag{8.18}$$

where r_t is the risk free rate, $\sigma_t \equiv \sigma_t^{\mathbb{F}}$ is the asset return volatility, $\theta_t^{\mathbb{F}}$ is the market price of $W^{\mathbb{F}}$-risk and $\left(r, \sigma, \theta^{\mathbb{F}}\right)$ are progressively measurable processes. The excess return process is $dR_t^e = \sigma_t \left(\theta_t^{\mathbb{F}} dt + dW_t^{\mathbb{F}}\right)$.

Private information consists of a signal G which is $\mathcal{F}_T \otimes \sigma(\varepsilon)$-measurable, where ε is an independent random variable representing noise. The informed filtration $\mathbb{G}$ is the enlargement of $\mathbb{F}$ by this private signal, $\mathbb{G} = \mathbb{F} \bigvee \sigma(G)$, where $\sigma(G)$ is the sigma-algebra generated by G.[9] The filtrations $\mathbb{F}$ and $\mathbb{G}$ are ordered: $\mathbb{G} \supseteq \mathbb{F}$. Excess returns have the representation $dR_t^e = \sigma_t^{\mathbb{G}} \left(\theta_t^{\mathbb{G}} dt + dW_t^{\mathbb{G}}\right)$ in $\mathbb{G}$.

In this setting, the PIPR is defined as follows,

Definition 3 *In the continuous time setting described above, where public information is carried by the filtration $\mathbb{F} = \{\mathcal{F}_t : t \in [0,T]\}$ generated by the Brownian motion $W^{\mathbb{F}}$, the PIPR associated with the augmented filtration $\mathbb{G} = \sigma(G) \vee \mathbb{F}$ is,*

$$\theta_t^{\mathbb{G},\mathbb{F}} = \lim_{h \to 0} \frac{1}{h} E\left[W_{t+h}^{\mathbb{F}} - W_t^{\mathbb{F}} \middle| \mathcal{G}_t \right] = \lim_{h \to 0} \frac{1}{h} E\left[\Delta W_{t,t+h}^{\mathbb{F}} \middle| \mathcal{G}_t \right] \tag{8.19}$$

for all $t \in [0,T]$. The excess return process is $dR_t^e = \sigma_t^{\mathbb{F}} \left(\left(\theta_t^{\mathbb{F}} + \theta_n^{\mathbb{G},\mathbb{F}} \right) dt + dW_t^{\mathbb{G},\mathbb{F}} \right)$ where $dW_t^{\mathbb{G},\mathbb{F}} = dW_t^{\mathbb{F}} - \theta_n^{\mathbb{G},\mathbb{F}} dt$ is a $\mathbb{G}$-Brownian motion. Moreover, $\sigma^{\mathbb{G}} = \sigma^{\mathbb{F}} = \sigma$ and $W^{\mathbb{G},\mathbb{F}} = W^{\mathbb{G}}$ so that $dR_t^e = \sigma_t \left(\left(\theta_t^{\mathbb{F}} + \theta_n^{\mathbb{G},\mathbb{F}} \right) dt + dW_t^{\mathbb{G}} \right)$.

The continuous time PIPR is the limit of the discrete time PIPR $E\left[\Delta W_{t,t+h}^{\mathbb{F}} \middle| \mathcal{G}_t \right]$ normalized by the length of the time interval h. It corresponds to the intensity of the compensator of the $\mathbb{F}$-return innovation $dW^{\mathbb{F}}$ in the private information $\mathbb{G}$: $\theta_t^{\mathbb{G},\mathbb{F}} dt \equiv E\left[dW_t^{\mathbb{F}} \middle| \mathcal{G}_t \right]$ for all $t \in [0,T]$.

There are two notable differences between the continuous time and discrete time models. In the continuous time Brownian setting, the quadratic variation and the volatility coefficient are defined independently of the filtration used for the Doob-Meyer decomposition. This follows because quadratic variation is a function of observations, $d\left[R^e\right]_t = d\left(R_t^e\right)^2 - 2R_t^e dR_t^e$. Expressing the excess return under each of the two filtrations gives $d\left[R^e\right]_t = \left(\sigma_t^{\mathbb{G}}\right)^2 dt = \left(\sigma_t^{\mathbb{F}}\right)^2 dt$, so that $\left|\sigma_t^{\mathbb{G}}\right| = \left|\sigma_t^{\mathbb{F}}\right|$. This holds for all $t \in [0,T]$, implying $\left|\sigma^{\mathbb{G}}\right| = \left|\sigma^{\mathbb{F}}\right|$. If $\sigma^{\mathbb{G}} = \sigma^{\mathbb{F}} = \sigma$, then $dW_t^{\mathbb{G},\mathbb{F}} - dW_t^{\mathbb{G}} = \left(\theta_t^{\mathbb{F}} + \theta_t^{\mathbb{G},\mathbb{F}} - \theta_t^{\mathbb{G}} \right) dt$ for $t \in [0,T]$, and, because $W^{\mathbb{G},\mathbb{F}}, W^{\mathbb{G}}$ are $\mathbb{G}$-Brownian motions, $E\left[dW_t^{\mathbb{G},\mathbb{F}} - dW_t^{\mathbb{G}} \middle| \mathcal{G}_t \right] = \left(\theta_t^{\mathbb{F}} + \theta_t^{\mathbb{G},\mathbb{F}} - \theta_t^{\mathbb{G}} \right) dt = 0$ for $t \in [0,T]$. Thus, $\theta_t^{\mathbb{G}} = \theta_t^{\mathbb{F}} + \theta_t^{\mathbb{G},\mathbb{F}}$ and $dW_t^{\mathbb{G},\mathbb{F}} = dW_t^{\mathbb{G}}$ for $t \in [0,T]$ (i.e., $W^{\mathbb{G},\mathbb{F}} = W^{\mathbb{G}}$). If $\sigma^{\mathbb{G}} = -\sigma^{\mathbb{F}} = -\sigma$, then $dW_t^{\mathbb{G},\mathbb{F}} + dW_t^{\mathbb{G}} = \left(\theta_t^{\mathbb{F}} + \theta_t^{\mathbb{G},\mathbb{F}} + \theta_t^{\mathbb{G}} \right) dt$ and $E\left[dW_t^{\mathbb{G},\mathbb{F}} + dW_t^{\mathbb{G}} \middle| \mathcal{G}_t \right] = \left(\theta_t^{\mathbb{F}} + \theta_t^{\mathbb{G},\mathbb{F}} + \theta_t^{\mathbb{G}} \right) dt = 0$ for $t \in [0,T]$. Thus, $\theta_t^{\mathbb{G}} = -\left(\theta_t^{\mathbb{F}} + \theta_t^{\mathbb{G},\mathbb{F}} \right)$ and $dW_t^{\mathbb{G},\mathbb{F}} = -dW_t^{\mathbb{G}}$

for $t \in [0,T]$. Redefining $dW_t^{\mathbb{G}} = -dW_t^{\mathbb{G}}$ and $\theta_t^{\mathbb{G}} = -\theta_t^{\mathbb{G}}$ leads to the same result as in the first case.

The relations in the continuous time Brownian model, between the volatilities and the Brownian motions under the filtrations $\mathbb{G}$ and $\mathbb{F}$, are the same as in a discrete time model with homoskedastic conditional variance (Proposition 2, (ii)). This property is intuitive. As the quadratic variation is an intrinsic property of a Brownian martingale, it does not change with the filtration.

The next proposition summarizes the properties of the PIPR in a continuous time model.

Proposition 4 (Properties of PIPR) *The PIPR has the following properties:*

(i) ***No-arbitrage:*** $E\left[\theta_t^{\mathbb{G},\mathbb{F}}\middle|\mathcal{F}_t\right] = 0$

(ii) ***Incremental price of risk:*** $\theta_t^{\mathbb{G},\mathbb{F}} = \theta_t^{\mathbb{G}} - \theta_t^{\mathbb{F}}$.

(iii) ***Covariation with log-density process:*** *If* $\mathcal{G}_t = \mathcal{F}_t \bigvee \sigma(G)$ *for some m-dimensional random vector* $G \in \mathcal{F}_N$,

$$\theta_t^{\mathbb{G},\mathbb{F}} = \frac{d\left[log\frac{P(G\in dx|\mathcal{F}_t)}{P(G\in dx)}, W^{\mathbb{F}}\right]_t}{dt}_{|x=G} \equiv \theta_t^{G|F}(x)_{|x=G} \tag{8.20}$$

(iv) ***Representation of density process of conditional signal:*** *The conditional probability density process has the representation,*

$$P(G \in dx|\mathcal{F}_t) = P(G \in dx)\,\mathcal{E}\left(\int_0^{\cdot} \theta_v^{G|F}(x)\,dW_v\right)_t \tag{8.21}$$

where $\mathcal{E}(M_\cdot)_t \equiv \exp\left(M_t - \frac{1}{2}[M]_t\right)$ *denotes the stochastic exponential. It follows that*

$$\theta_t^{G|F}(x) = \mathcal{D}_t log\frac{P(G \in dx|\mathcal{F}_t)}{P(G \in dx)} \tag{8.22}$$

where $\mathcal{D}_t$ *denotes the Malliavin derivative operator.*

The interpretation of the no-arbitrage property $E\left[\theta_t^{\mathbb{G},\mathbb{F}}\middle|\mathcal{F}_t\right] = 0$ and of the PIPR as a compensation for incremental risk is the same as in the discrete time model. The representation property in (iv) follows from the $\mathbb{F}$-martingale property of the conditional measure $P(G \in dx|\mathcal{F}_t)$ and the Clark-Ocone formula (see the proof in the appendix for details). The covariation representation in property (iii) is an immediate consequence of (iv).

It is also interesting to note that the PIPR is related to the conditional probability of the fundamental source of uncertainty $W_\cdot^{\mathbb{F}}$. To see this, use Bayes' rule to write,

$$\frac{P\left(W_\cdot^{\mathbb{F}} \in d\omega\middle|G = x\right)}{P\left(W_\cdot^{\mathbb{F}} \in d\omega\right)} = \frac{P(G \in dx|\mathcal{F}_t)}{P(G \in dx)} = \mathcal{E}\left(\int_0^{\cdot} \theta_v^{G|F}(x)\,dW_v^{\mathbb{F}}\right) \tag{8.23}$$

where the second equality follows from the representation property (iv). The conditional density process $\mathcal{E}\left(\int_0^{\cdot} \theta_v^{G|F}(x)\, dW_v^{\mathbb{F}}\right)$, therefore defines a conditional Wiener measure. That is, it defines the distribution of $W^{\mathbb{F}}$ conditional on the realization $G = x$, the numerator on the left hand side of (8.23). Viewed from this perspective, the change of information from $\mathbb{F}$ to $\mathbb{G}$, is closely related to a change in beliefs. The Brownian motion $W_{\cdot}^{\mathbb{G},\mathbb{F}}$ is a Brownian motion under the conditional Wiener measure evaluated at the true realization of the signal $x = G$. The PIPR is therefore determined by the disagreement between the conditional and the unconditional Wiener measures. When the two agree, information has no value and $\theta^{G|F}(x) = 0$.

The notion that the PIPR quantifies the information content of the signal can be strengthened by looking at entropy. Define the conditional Wiener measure $P^x(d\omega) \equiv P(d\omega|\, G = x)$, i.e., the measure on Brownian paths conditional on $G = x$. The relative entropy, or Kullback-Leibler (KL) divergence, of the unconditional Wiener measure with respect to the conditional Wiener measure is,

$$D_{KL}\left(P||P^x\right) \equiv E\left[\log \frac{dP}{dP^x}\right] = \frac{1}{2} E\left[\int_0^T \|\theta_v^{G|F}(G)\|^2 dv\right].$$

The divergence measure $D_{KL}(P||P^x)$, also called the information gain, quantifies the discrepancy between the unconditional and conditional measures.[10] It is fully determined by the PIPR. Moreover, given a signal realization, divergence is large if and only if the absolute value of the PIPR is large. This property is natural given the interpretation of the PIPR as the incremental price of risk associated with the private information signal.

3.2 Examples

3.2.1 Example 1: Noisy level signal

Consider the price model (8.18) with constant coefficients $\left(r, \sigma, \theta^{\mathbb{F}}\right)$. Suppose that the private signal $G = S_{T-h,T}\zeta$ with $S_{T-h,T} = S_T / S_{T-h}$ and with $\zeta = \exp\left(\sigma_\zeta W_1^\zeta - \frac{1}{2}\sigma_\zeta^2\right)$ an independent random variable such that $E[\zeta] = 1$, $E\left[\log \zeta\right] = -\frac{1}{2}\sigma_\zeta^2$ and $VAR\left[\log \zeta\right] = \sigma_\zeta^2$. The signal G provides noisy information about the growth rate of the price between $T-h$ and T. Equivalently, it provides noisy information about the level of the terminal price. To simplify notation, let $E[\cdot|\mathcal{F}_v] \equiv E_v[\cdot]$. The conditional density of the signal, for $v \in [T-h, T)$, given $Y_{\tau_i} = G_i$ is,

$$p_v^G(x) \equiv \partial_x P_v(G \le x) = \partial_x P_v\left(\log\left(S_{T-h,T}\right) + \sigma_\zeta W_1^\zeta \le \log(x) + \frac{1}{2}\left(\sigma_\zeta^2\right)\right)$$

$$= \partial_x P_v \left(\frac{\log\left(S_{v,T}\right) - E_v\left[\log\left(S_{v,T}\right)\right] + \sigma_\zeta W_1^\zeta}{\sqrt{VAR_v\left[\log G\right]}} \leq d\left(x,v\right) \right)$$

$$= \partial_x \Phi \left(\frac{\log x - E_v\left[\log G\right]}{\sqrt{VAR_v\left[\log G\right]}} \right) = \frac{1}{x\sqrt{VAR_v\left[\log G\right]}} \phi\left(d\left(x,v\right)\right)$$

where $d\left(x,v\right) \equiv \left(\log x - E_v\left[\log G\right]\right)/\sqrt{VAR_v\left[\log G\right]}$ and use was made of,

$$E_v\left[\log\left(S_{T-h,T}\right)\right] = \log S_{T-h,v} + \left(r + \sigma\theta^{\mathbb{F}} - \frac{1}{2}\sigma^2\right)\left(T-v\right)$$

$$\log\left(S_{T-h,T}\right) - E_v\left[\log\left(S_{T-h,T}\right)\right] = \log\left(S_{v,T}\right) - E_v\left[\log\left(S_{v,T}\right)\right]$$

$$E_v\left[\log G\right] = E_v\left[\log\left(S_{T-h,T}\right)\right] - \frac{1}{2}\sigma_\zeta^2, \qquad VAR_v\left[\log G\right] = \sigma^2\left(T-v\right) + \sigma_\zeta^2.$$

As $\theta_v^{G|F}\left(x\right) = d\left[\log p_\cdot(x), W_\cdot^{\mathbb{F}}\right]_v/dv$ (Proposition 4 (iii)) it follows that,

$$\theta_v^{G|F}\left(x\right) = \frac{d\left[p_\cdot^G(x), W_\cdot^{\mathbb{F}}\right]_v}{p_v^G(x)dv} = \frac{-\phi'\left(d\left(x,v\right)\right)}{\phi\left(d\left(x,v\right)\right)\sqrt{VAR_v\left[\log G\right]}} \frac{d\left[\log S_{T-h,\cdot}, W_\cdot^{\mathbb{F}}\right]_v}{dv}$$

$$= \sigma\left(\frac{\log x - E_v\left[\log G\right]}{VAR_v\left[\log G\right]}\right)$$

(because $\phi'(x) = -x\phi\left(x\right)$ and $d\left[\log S_{T-h,\cdot}, W_\cdot^{\mathbb{F}}\right]_v = \sigma dv$) and

$$\theta_v^{\mathbb{G},\mathbb{F}} \equiv \theta_v^{G|F}\left(G\right) = \sigma\left(\frac{\log G - E_v\left[\log G\right]}{VAR_v\left[\log G\right]}\right). \tag{8.24}$$

Formula (8.24) is the continuous time, noisy version of the PIPR in the discrete time model with Gaussian innovations. The sign of the PIPR depends on the scaled innovation $\frac{\log G - E_v[\log G]}{STD_v[\log G]}$. It is large when the scaled return innovation and/or the ratio $\sigma/STD_v\left[\log G\right]$ is large. As $VAR_v\left[\log G\right] = \sigma^2\left(T-v\right) + \sigma_\zeta^2$, the PIPR explodes in the absence of noise ($\sigma_\zeta^2 = 0$) as time T approaches and uncertainty about the signal is resolved. Without noise, information becomes so valuable that the PIPR goes to $\pm\infty$. This behavior reflects the emergence of an arbitrage opportunity in the limit.[11]

This property is also reflected by the KL divergence measure of the signal. The relative entropy (see Detemple and Rindisbacher (2013) for a proof) is,

$$D_{KL}\left(P||P^G\right) = \sqrt{1 + \left(\frac{\sigma\sqrt{h}}{\sigma_\zeta}\right)^2} \tag{8.25}$$

Relative entropy is deterministic and large if the signal-to-noise ratio σ/σ_ζ is large, i.e., if the volatility of the stock price is high and/or if the standard deviation of the signal noise is small.

3.2.2 Example 2: Noisy directional signal

Consider a setting where the price process is the same as in Example 1, but the noisy signal is a directional signal given by,

$$G = g\left(S_T, \zeta_i\right) = sign\left(\log S_{T-h,T}\right) sign\left(\zeta - F_\zeta^{-1}(p)\right) \tag{8.26}$$

where $sign(x) = 1_{\{x>0\}} - 1_{\{x\leq 0\}}$ is the sign of x and ζ is an independent noise component with survival function $F_\zeta(x) = P(\zeta > x)$. The signal G indicates the correct direction of the stock price with probability p. A skilled investor is an investor with $p > 1/2$. Perfect directional forecasting ability corresponds to the case $p = 1$.

Let $P_v(E) \equiv P(E|\mathcal{F}_v)$ and $S_{T-h,T} = S_T/S_{T-h}$. The conditional distribution of the signal for $v \in [T-h, T)$ is,

$$\begin{aligned}
p_v^G(y) \equiv P_v(G=y) &= \left(P_v\left(\log S_{T-h,T} > 0\right)p + P_v\left(\log S_{T-h,T} \leq 0\right)(1-p)\right) 1_{y=1} \\
&\quad + \left(P_v\left(\log S_{T-h,T} > 0\right)(1-p) + P_v\left(\log S_{T-h,T} \leq 0\right)p\right) 1_{y=-1} \\
&= \left(p + (1-2p) P_v\left(\log S_{T-h,T_i} \leq 0\right)\right) 1_{y=1} \\
&\quad + \left(1 - p + (2p-1) P_v\left(\log S_{T-h,T} \leq 0\right)\right) 1_{y=-1} \\
&= \left(p 1_{y=1} + (1-p) 1_{y=-1}\right) + \left((1-2p) 1_{y=1} + (2p-1) 1_{y=-1}\right) \Phi\left(d\left(S_{T-h,v}, v\right)\right) \\
&= \left(p 1_{y=1} + (1-p) 1_{y=-1}\right) + (1-2p)\left(1_{y=1} - 1_{y=-1}\right) \Phi\left(d\left(S_{T-h,v}, v\right)\right) \\
&= \left(p 1_{y=1} + (1-p) 1_{y=-1}\right) + (1-2p) y \Phi\left(d\left(S_{T-h,v}, v\right)\right) \quad \text{for} \quad y \in \{-1, 1\}
\end{aligned}$$

where

$$d\left(S_{T-h,v}, v\right) \equiv \frac{-E_v\left[\log S_{T-h,T-h}\right]}{\sqrt{VAR_v\left[\log S_{T-h,T}\right]}} = \frac{-\log S_{T-h,v} - E_v\left[\log S_{v,T}\right]}{\sqrt{VAR_v\left[\log S_{v,T}\right]}}.$$

Using $d\left[\Phi\left(d\left(S_{T-h,\cdot}, v\right)\right), W_\cdot^{\mathbb{F}}\right]_v / dv = -\phi\left(d\left(S_{T-h,v}, v\right)\right)/\sqrt{T-v} \equiv -g\left(S_{T-h,v}, v\right)$ and $d\left[\log X_\cdot, W_\cdot\right]_v = d\left[X_\cdot, W_\cdot^{\mathbb{F}}\right]_v / X_v$, gives,

$$\theta_v^{\mathbb{G},\mathbb{F}} = \theta_v^G(G) = \left.\frac{d\left[\log p_\cdot^G(x), W_\cdot\right]_v}{dv}\right|_{x=G} = 2sg\left(S_{T-h,v}, v\right) H_v(G)$$

where $s \equiv p - 1/2$ measures skill and where

$$H_v(y) \equiv y p_v^G(y)^{-1} = y\left[\frac{1}{2} + s\left(2y\left(1 - \Phi\left(d\left(S_{T_{i-1},v}, v\right)\right)\right) + 1_{y=-1} - 1_{y=1}\right)\right]^{-1}.$$

For the directional signal, the KL divergence measure becomes (see Detemple and Rindisbacher (2013) for a proof),

$$D_{KL}\left(P || P^G\right) = \exp\left(p \log p + (1-p)\log(1-p) - \sum_{z \in \{-1,1\}} \log p_{T-h}^G(z) p_{T-h}^G(z)\right)$$

where the conditional density of the signal is,

$$p^G_{T-h}(z) \equiv 2(1-\Phi(d(1,T-h)))\,s\,(\mathbf{1}_{z=1}-\mathbf{1}_{z=-1}) + (1-p)\,\mathbf{1}_{z=1} + p\mathbf{1}_{z=-1}$$

for $z \in \{-1,1\}$. The value of information is seen to be high if the probability of making a correct forecast p and therefore the skill level $s = p - 1/2$ is high.

3.3 Optimal informed portfolio and PIPR

Consider an investor with information $\mathbb{G}$ and logarithmic utility $U_0 = E\left[\log X_T | \mathcal{G}_0\right]$. Letting $b_T \equiv \exp\left(\int_0^T r_s ds\right)$, it is easy to verify that discounted terminal wealth satisfies,

$$\frac{X_T}{b_T} = X_0 + \int_0^T \frac{X_s}{b_s} \pi_s' \sigma_s \left(\theta_s^{\mathbb{G}} ds + dW_s^{\mathbb{G},\mathbb{F}}\right)$$

and that,

$$\log X_T = \log X_0 + \log b_T + \int_0^T \pi_s' \sigma_s \left(\theta_s^{\mathbb{G}} ds + dW_s^{\mathbb{G},\mathbb{F}}\right) - \frac{1}{2}\int_0^T \|\pi_s'\sigma_s\|^2 ds.$$

It follows that,

$$U_0^{\mathbb{G}} = \log X_0 + \log b_T + \frac{1}{2} E\left[\int_0^T \|\theta_s^{\mathbb{G}}\|^2 ds \middle| \mathcal{G}_0\right] - \frac{1}{2} E\left[\int_0^T \|\sigma_s'\pi_s - \theta_s^{\mathbb{G}}\|^2 ds \middle| \mathcal{G}_0\right].$$

The optimal portfolio of the logarithmic investor is the one that minimizes the last term of $U_0^{\mathbb{G}}$, that is,

$$\pi_t^* = \left(\sigma_t'\right)^{-1} \theta_t^{\mathbb{G}} = \left(\sigma_t'\right)^{-1} \theta_t^{\mathbb{F}} + \left(\sigma_t'\right)^{-1} \theta_t^{\mathbb{G},\mathbb{F}}.$$

The optimal portfolio has a mean-variance structure with two components. The first component, $\pi_t^{\mathbb{F},*} \equiv \left(\sigma_t'\right)^{-1} \theta_t^{\mathbb{F}}$, exploits all diversification benefits given public information. Optimal private information trading is captured by the second component, $\pi_t^{\mathbb{G},*} \equiv \left(\sigma_t'\right)^{-1} \theta_t^{\mathbb{G},\mathbb{F}}$. Given the no-arbitrage property, $E\left[\theta_t^{\mathbb{G},\mathbb{F}} \middle| \mathcal{F}_t\right] = 0$, information trading perceived by the public vanishes on average, $E\left[\pi_t^{\mathbb{G}} | \mathcal{F}_t\right] = 0$.

The optimal utility of the informed investor with private information $\mathbb{G}$ can be expressed in terms of the utility of an uninformed investor with same preferences and initial wealth, but using public information $\mathbb{F}$ for decision making. Specifically,

$$U_0^{\mathbb{G},*} = E\left[\log X_0 + \log b_T\right] + \frac{1}{2} E\left[\int_0^T \|\theta_s^{\mathbb{G}}\|^2 ds\right] = U_0^{\mathbb{F},*} + D_{KL}\left(P || P^G\right).$$

where $U_0^{\mathbb{F},*} = E\left[\log X_0 + \log b_T\right] + \frac{1}{2} E\left[\int_0^T \|\theta_s^{\mathbb{F}}\|^2 ds\right]$ is the optimal uniformed utility and $U_0^{\mathbb{G},*}$ the optimal informed utility. The welfare difference between the two agents is the KL divergence measure. This welfare decomposition result depends crucially on the no-arbitrage property of the PIPR. It shows that the utility gain of an informed log-investor is large if and only if the relative entropy of the signal is large. An informed investor who has a mutually exclusive choice between J different signals $\{G_j : j = 1, \ldots, J\}$, will optimally choose the signal with the largest information gain as measured by the relative entropy measure. The optimal signal is given by, $G_{j^*} = \arg\sup_{\{G_j : j=1,\ldots,J\}} D_{KL}\left(P || P^{G_j}\right)$.[12]

3.4 Testing for skill

Given the additive structure of the optimal portfolio $\pi_t^* = \pi_t^{\mathbb{F},*} + \pi_t^{\mathbb{G},*}$ where $\pi_t^{\mathbb{F},*} = (\sigma_t')^{-1}\theta_t^{\mathbb{F}}$ and $\pi_t^{\mathbb{G},*} = (\sigma_t')^{-1}\theta_t^{\mathbb{G},\mathbb{F}}$, the expected excess return of an informed trader, conditional on public information, is given by,

$$E\left[\left.\frac{dX_t}{X_t} - r_t dt\right|\mathcal{F}_t\right] = E\left[\left.\|\theta_t^{\mathbb{G}}\|^2\right|\mathcal{F}_t\right]dt = \|\theta_t^{\mathbb{F}}\|^2 dt + E\left[\left.\|\theta_t^{\mathbb{G},\mathbb{F}}\|^2\right|\mathcal{F}_t\right]dt$$

where we used the fact that $\|\theta_t^{\mathbb{F}}\|^2$ is $\mathcal{F}_t$-measurable and where the first component $\|\theta_t^{\mathbb{F}}\|^2 dt$ represents the expected excess return achieved on the basis of public information. It follows, from this expression, that an informed investor is perceived by the public to be skilled if and only if $\theta_t^{\mathbb{G},\mathbb{F}} \neq 0$. If $\theta_t^{\mathbb{G},\mathbb{F}} \neq 0$, the incremental expected excess return generated by informed trading is $E\left[\left.\|\theta_t^{\mathbb{G},\mathbb{F}}\|^2\right|\mathcal{F}_t\right]dt > 0$.

The presence of skill associated with private information introduces the timing factor $E\left[\left.\|\theta_t^{\mathbb{G},\mathbb{F}}\|^2\right|\mathcal{F}_t\right]dt$ in the regression representation of the portfolio excess return. Tests for the presence of skill can therefore be implemented as tests of significance of this timing factor. The structural market timing model developed in Detemple and Rindisbacher (2013) derives the timing factor from the optimal trading strategy of the informed investor. Their analysis extends the timing regression models for skill found in the earlier literature, such as the level forecast model pioneered by Treynor and Mazuy (1966) and the directional model introduced by Merton (1981) and Henriksson and Merton (1981). In these classical studies, time is discrete and trading occurs at the observation (reporting) frequency. Goetzmann, Ivkovic and Ingersoll (2002) identify non-structural timing factors for situations where trading occurs at a higher frequency than the reporting frequency of fund returns. Their analysis permits the correction of biases resulting from the frequency mismatch. Both the Treynor-Mazuy level forecast model and the Henriksson-Merton directional model are fully parametric, therefore potentially misspecified. To address the misspecification of parametric timing models, Glosten and Jagannathan (1984) formulate a semi-parametric model for timing factors. Breen, Jagannathan and Ofer (1986) emphasize the fact that timing regressions typically have a heteroskedastic error structure. Factor structures for the market component $\|\theta_t^{\mathbb{F}}\|$, in models of hedge fund returns, are discussed by Brown and Goetzmann (2003), Fung and Hsieh (2001) and Chan et al. (2005). Empirical timing regressions are applied to mutual funds and hedge funds by Henriksson (1984), Ferson and Schadt (1996), Ferson and Khang (2002), Ferson et al. (2006), Chen (2007) and Chen and Liang (2007). Admati et al. (1986) and Detemple and Rindisbacher (2013) propose approaches to distinguish between timing skill and selection skill. Performance measurement, based on the CAPM, in the presence of asymmetric information is examined by Dybvig and Ross (2005). Detemple and Rindisbacher (2013) show that the PIPR is the key quantity required to analyze the measurement of skill and the performance of skilled fund managers.

4 Conclusion

The PIPR quantifies the incremental price of risk derived from the use of private information relative to public information. Knowledge of the PIPR helps to identify the risk-reward trade-off faced by an informed investor. It lies at the heart of the optimal portfolio of the informed and ultimately determines the welfare gains associated with the collection of private information. It is a critical factor underlying the performance of portfolio managers and the object of interest for performance measurement and tests of skill.

The identification of the PIPR in consumption-portfolio models with private information is relatively straightforward. As explained, the PIPR can be calculated from the distributional properties of the asset return and the private signal collected. Its determination in equilibrium settings is more challenging. In these models, private information can leak in the market and be reflected, to some extent, in prices and residual demands. Some initially private information effectively becomes public. In these cases, the PIPR must be calculated relative to an endogenous public filtration. As the PIPR determines demand functions and ultimately prices, this a priori unknown filtration itself depends on the PIPR. The resolution of this fixed point problem is non trivial. Detemple, Rindisbacher and Truong (2014) solve the problem in a dynamic model with private information about the terminal dividend payment of a stock. They show that the PIPR becomes a function of a noisy statistic, revealed by the equilibrium price, for the private signal of the informed. They study the local value of the private signal, the trading behavior and welfare of the agents and the equilibrium price properties.

Notes

Questrom School of Business, Boston University, 595 Commonwealth Ave Boston, MA 02215. Jerome Detemple: detemple@bu.edu, Marcel Rindisbacher: rindisbm@bu.edu

1. A fundamental source of risk is understood to be a return innovation. Such a random variable has zero mean and unit variance. The stochastic process formed by cumulating return innovations is a martingale.
2. A process x is adapted with respect to $\mathbb{F}$ if and only if $x_n = x(n,\omega)$ is $\mathcal{F}_n$-measurable for each $n \in \mathbb{N}$.
3. $\mathcal{F}_n \bigvee \sigma(G)$ is the enlargement of $\mathcal{F}_n$ by the information conveyed by the signal G.

4. In the continuous time limit, the density process is a stochastic exponential. It is easy to see this when the market is complete. A Taylor expansion then gives, $Z^G_{n,n+1} = \exp\left(\log\left(1+\theta^{G|F}_n(x)\,\Delta W^{\mathbb{F}}_n\right)\right) = \exp\left(\theta^{G|F}_n(x)\,\Delta W^{\mathbb{F}}_n - \frac{1}{2}\theta^{G|F}_n(x)^2\left(\Delta W^{\mathbb{F}}_n\right)^2 + o_P\left(\left(\Delta W^{\mathbb{F}}_n\right)^2\right)\right)$. Under appropriate scaling, by Donsker's invariance principle (see Karatzas and Shreve (1988)), $\Delta W^{\mathbb{F}}_n \to dW^{\mathbb{F}}_t$ in probability when the length of each time interval goes to 0, where $W^{\mathbb{F}}$ is a $(P,\mathbb{F})$-Browian motion. Hence, $\left(\Delta W^{\mathbb{F}}_n\right)^2 \to dt$ in probability and the limit density is $Z^G_{t,s} = \exp\left(\int_t^s \theta^{G|F}_v(x)\,dW^{\mathbb{F}}_v - \frac{1}{2}\int_t^s \theta^{G|F}_v(x)^2\,dv\right)$. This expression implies that $\theta^{G|F}_s(x)\,ds$ is the covariation at time s between the density process and the return innovation under public information.
5. An atom of a σ-algebra $\mathcal{F}$ is an event $A \in \mathcal{F}$ such that, for any $E \in \mathcal{F}$, $E \subset A$ implies $E = A$ or $E = \emptyset$.
6. Suppose that the length of each period is h, instead of 1. If the PIPR is redefined as a rate per unit time, then $\theta^{\mathbb{G},\mathbb{F}}_1 h = \Delta W^{\mathbb{F}}_1$. Assuming that $\Delta W^{\mathbb{F}}_1 = w^{\mathbb{F}}_1\sqrt{h}$, it follows that $\theta^{\mathbb{G},\mathbb{F}}_1 \to \pm\infty$ as $h \to 0$. This previews some of the properties in the continuous time setting.
7. Exact conditions for the absence of arbitrage opportunities and free lunches with vanishing risk in continuous time are discussed in Rindisbacher (1999), Imkeller et al. (2001) and Imkeller (2003).
8. Suppose $d = 1$. If the private signal is fully revealing, $\kappa_0(\tau) = 0$ and the portfolio problem becomes degenerate, $\max_{\pi_0 \in \mathcal{G}_0}\left\{A\pi_0\sigma^{\mathbb{F}}_0\left(\theta^{\mathbb{F}}_0 + \theta^{\mathbb{G},\mathbb{F}}_0\right)\right\}$. The solution is $\pm\infty$ depending on the sign of $\sigma^{\mathbb{F}}_0\left(\theta^{\mathbb{F}}_0 + \theta^{\mathbb{G},\mathbb{F}}_0\right)$.
9. $\mathbb{F}\bigvee\sigma(G) = \left\{\mathcal{F}_t\bigvee\sigma(G) : t \in [0,T]\right\}$ where $\mathcal{F}_t\bigvee\sigma(G)$ is the sigma-algebra generated by $\sigma(G)$ and $\mathcal{F}_t$.
10. Consider a random variable Z with distribution $P(Z \in dz) = p(z)\,dz$. The Shannon entropy of Z, defined as $H(Z) \equiv -E\left[\log p(Z)\right]$, measures the amount of uncertainty (or disorder) associated with Z (Shannon (1948)).
11. See Rindisbacher (1999), Imkeller et al. (2001) and Imkeller (2003) for the relation between explosions of the PIPR, singularities, free lunches with vanishing risk and arbitrage opportunities.
12. Ankirchner et al. (2005) relate the additional utility of the informed to the Shannon information of the filtration.

References

[1] Admati, A.R., S. Battacharya, P. Pfleiderer and S.A. Ross, 1986, "On Timing and Selectivity," *Journal of Finance* 41, 735–730.

[2] Amendinger, J., P. Imkeller and M. Schweizer, 1998, "Additional Logarithmic Utility of an Insider," *Stochastic Processes and their Applications* 75, 263–286.

[3] Amendinger, J., D. Becherer and M. Schweizer, 2003, "A Monetary Value for Initial Information in Portfolio Optimization," *Finance and Stochastics* 7, 29–46.

[4] Ankirchner, S., and P. Imkeller, 2005, "Finite Utility on Financial Markets with Asymmetric Information and Structure Properties of the Price Dynamics," *Annales de l'Institut Henri Poincaré: Probabilité et Statistique* 41, 479–503.

[5] Ankirchner, S., P.S. Dereich and P. Imkeller, 2006, "The Shannon Information of Filtrations and the Additional Logarithmic Utility of Insiders," *Annals of Probability* 34, 743–778.

[6] Biagini, F., and B. Oksendal, 2005, “A General Stochastic Calculus Approach to Insider Trading,” *Applied Mathematics and Optimization* 52, 167–181.

[7] Breen, W., R. Jagannathan and A.R. Ofer, 1986, “Correcting for Heteroscedasticity in Tests for Market Timing,” *Journal of Business* 59, 585–598.

[8] Brown, S.J., and W. Goetzmann, 2003, “Hedge Funds with Style,” *Journal of Portfolio Management* 29, 101–112.

[9] Chan, N., M. Getmansky, S.M. Haas and A. Lo, 2005, “Systemic Risk and Hedge Funds,” Working Paper, Massachusetts Institute of Technology.

[10] Chen, Y., 2007, “Timing Ability in the Focus Market of Hedge Funds,” *Journal of Investment Management* 5, 66–98.

[11] Chen, Y., and B. Liang, 2007, “Do Market Timing Hedge Funds Time the Market?,” *Journal of Financial and Quantitative Analysis* 42, 827–856.

[12] Detemple, J., and M. Rindisbacher, 2013, “A Structural Model of Dynamic Market Timing,” *Review of Financial Studies* 16, 2492–2547.

[13] Detemple, J., M. Rindisbacher and T. Truong, 2014, “Dynamic Noisy Rational Expectations Equilibria with Anticipative Information,” Working Paper, Boston University.

[14] Dybvig, P.H., and S.A. Ross, 1985, “Differential Information and Performance Measurement Using a Security Market Line,” *Journal of Finance* 40, 383–399.

[15] Ferson, W., T.R. Henry and D.J. Kisgen, 2006, “Evaluating Government Bond Fund Performance with Stochastic Discount Factors,” *Review of Financial Studies* 19, 423–455.

[16] Ferson, W., and K. Khang, 2002, “Conditional Performance Measurement using Portfolio Weights: Evidence for Pension Funds,” *Journal of Financial Economics* 65, 249–282.

[17] Ferson, W.E., and R.W. Schadt, 1996, “Measuring Fund Strategy Performance in Changing Economic Conditions,” *Journal of Finance* 51, 425–461.

[18] Fung, W., and D.A. Hsieh, 2001, “The Risk in Hedge Fund Strategies: Theory and Evidence from Trend Followers,” *Review of Financial Studies* 14, 313–341.

[19] Glosten, L.R., and R. Jagannathan, 1994, “A Contingent Claim Approach to Performance Evaluation,” *Journal of Empirical Finance* 1, 133–160.

[20] Goetzmann, W., J. Ingersoll and I. Ivkovic, 2002, “Monthly Measurement of Daily Timers,” *Journal of Financial and Quantitative Analysis* 35, 257–290.

[21] Grorud, A., and M. Pontier, 1998, “Insider Trading in a Continuous Time Market Model,” *International Journal of Theoretical and Applied Finance* 1, 331–347.

[22] Grossman, S.J., and J.E. Stiglitz, 1980, “On the Impossibility of Informationally Efficient Markets,” *American Economic Review* 70, 393–408.

[23] Henriksson, R.D., 1984, “Market Timing and Mutual Fund Performance: An Empirical Investigation,” *Journal of Business* 57, 73–96.

[24] Henriksson, R.D., and R.C. Merton, 1981, “On Market Timing and Investment Performance II: Statistical Procedures for Evaluating Forecasting Skills,” *Journal of Business* 54, 513–533.

[25] Imkeller P., 1996, “Enlargement of Wiener Filtration by an Absolutely Continuous Random Variable via Malliavin's Calculus,” *Probability Theory and Related Fields* 106, 105–135.

[26] Imkeller, P., 2003, “Malliavin's Calculus in Insider Models: Additional Utility and Free Lunches,” *Mathematical Finance* 13, 153–169.

[27] Imkeller, P., M. Pontier and F. Weisz, 2001, "Free Lunch and Arbitrage Possibilities in a Financial Market Model with an Insider," *Stochastic Processes and their Application* 92, 103–130.

[28] Jacod, J., "Grossissement Initial, Hypothèse (H) et Théorème de Girsanov," in T. Jeulin and M. Yor (Eds.), *Grossissements de Filtrations: Exemples et Applications*, Lecture Notes in Mathematics 1118, Springer, Berlin, 1985.

[29] Jeulin, T., *"Semi-Martingales et Grossissement de Filtration,"* Lecture Notes in Mathematics 833, Springer, Berlin, 1980.

[30] Karatzas, I., and I. Pikovsky, 1996, "Anticipative Portfolio Optimization," *Advances in Applied Probability* 28, 1095–1122.

[31] Karatzas, I., and S. Shreve, 1988, *Brownian Motion and Stochastic Calculus*, Springer-Verlag.

[32] Merton, R.C., 1981, "On Market Timing and Investment Performance I: An Equilibrium Theory of Value for Market Forecasts," *Journal of Business* 54, 363–406.

[33] Rindisbacher, M., 1999, "Insider Information, Arbitrage, and Optimal Consumption and Investment Policies," Working Paper, Université de Montréal.

[34] Shannon, C.E., 1948, "A Mathematical Theory of Communication," *Bell System Technical Journal* 27, 379–423, 623-656.

[35] Treynor, J., and K. Mazuy, 1966, "Can Mutual Funds Outguess the Market?" *Harvard Business Review* 44, 131–136.

5 Appendix: Proofs

Proof of Proposition 2.

(i) By definition of the PIPR, the tower property of conditional expectation, and the fact that $\Delta W_n^{\mathbb{F}}$ is an $\mathbb{F}$-martingale difference sequence, $E\left[\theta_n^{\mathbb{G}} \middle| \mathcal{F}_n\right] = E\left[E\left[\Delta W_n^{\mathbb{F}} \middle| \mathcal{G}_n\right] \middle| \mathcal{F}_n\right] = E\left[\Delta W_n^{\mathbb{F}} \middle| \mathcal{F}_n\right] = 0$.

(ii) As $\sigma_n^{\mathbb{G}}\left(\theta_n^{\mathbb{G}} + \Delta W_n^{\mathbb{G}}\right) = \sigma_n^{\mathbb{F}}\left(\theta_n^{\mathbb{F}} + \theta_n^{\mathbb{G},\mathbb{F}} + \Delta W_n^{\mathbb{G},\mathbb{F}}\right)$, taking conditional expectations and using the fact that $E\left[\Delta W_n^{\mathbb{G}} \middle| \mathcal{G}_n\right] = E\left[\Delta W_n^{\mathbb{G},\mathbb{F}} \middle| \mathcal{G}_n\right] = 0$ gives $\sigma_n^{\mathbb{G}}\theta_n^{\mathbb{G}} = \sigma_n^{\mathbb{F}}\left(\theta_n^{\mathbb{F}} + \theta_n^{\mathbb{G},\mathbb{F}}\right)$. Solving for the PIPR establishes the result.

(iii) Given the structure of the signal,

$$
\begin{aligned}
E\left[\Delta W_n^{\mathbb{F}} \middle| \mathcal{G}_n\right] &= \int_{\mathbb{R}} wP\left(\Delta W_n^{\mathbb{F}} \in dw \middle| G = x, \mathcal{F}_n\right)_{|x=G} \\
&= \int_{\mathbb{R}} w\left(\frac{P\left(\Delta W_n^{\mathbb{F}} \in dw, G \in dx \middle| \mathcal{F}_n\right)}{P(G \in dx | \mathcal{F}_n)}\right)_{|x=G} \\
&= \int_{\mathbb{R}} w\frac{P\left(G \in dx | \Delta W_n^{\mathbb{F}} = w, \mathcal{F}_n\right)}{P(G \in dx | \mathcal{F}_n)} P\left(\Delta W_n^{\mathbb{F}} \in dw \middle| \mathcal{F}_n\right)_{|x=G} \\
&= \int_{\mathbb{R}} w\frac{E\left[P(G \in dx | \mathcal{F}_{n+1}) | \Delta W_n^{\mathbb{F}} = w, \mathcal{F}_n\right]}{P(G \in dx | \mathcal{F}_n)} P\left(\Delta W_n^{\mathbb{F}} \in dw \middle| \mathcal{F}_n\right)_{|x=G} \\
&= E\left[\Delta W_n^{\mathbb{F}} \frac{P(G \in dx | \mathcal{F}_{n+1})}{P(G \in dx | \mathcal{F}_n)} \middle| \mathcal{F}_n\right]_{|x=G},
\end{aligned}
$$

and, as by definition $\theta_n^{\mathbb{G},\mathbb{F}} = E\left[\Delta W_n^{\mathbb{F}} \mid \mathcal{G}_n\right]$, the result follows.

(iv) Let $\Delta V_n^{\mathbb{F}}(x) = Z_{n,n+1}^{G}(x) - 1 - \theta_n^{G|F}(x)' \Delta W_n^{\mathbb{F}}$. Straightforward calculations give $E\left[\Delta V_n^{\mathbb{F}}(x) \mid \mathcal{F}_n\right] = E\left[Z_{n,n+1}^{G}(x) \middle| \mathcal{F}_n\right] -1 \; -\theta_n^{G|F}(x)' E\left[\Delta W_n^{\mathbb{F}} \mid \mathcal{F}_n\right] = 1 - 1 + 0 = 0$ and $E\left[\Delta V_n^{\mathbb{F}}(x)\left(\Delta W_n^{\mathbb{F}}\right)' \middle| \mathcal{F}_n\right] = E\left[Z_{n,n+1}^{G}(x)\left(\Delta W_n^{\mathbb{F}}\right)' \middle| \mathcal{F}_n\right] - \theta_n^{G|F}(x)' = 0'$. This establishes the representation of the conditional density process of the signal.

■

Proof of Proposition 4.

(iv) The process $P(G \in dx \mid \mathcal{F}_t)$ is a martingale. By the Clark-Ocone formula,

$$\begin{aligned} P(G \in dx \mid \mathcal{F}_t) &= E[P(G \in dx \mid \mathcal{F}_t)] + \int_0^t E[\mathcal{D}_v P(G \in dx \mid \mathcal{F}_t) \mid \mathcal{F}_v]\, dW_v^{\mathbb{F}} \\ &= P(G \in dx) + \int_0^t P(G \in dx \mid \mathcal{F}_v)\, \mathcal{D}_v \log E[P(G \in dx \mid \mathcal{F}_t) \mid \mathcal{F}_v]\, dW_v^{\mathbb{F}} \\ &= P(G \in dx) + \int_0^t P(G \in dx \mid \mathcal{F}_v)\, \theta_v^{G|F}(x)\, dW_v^{\mathbb{F}}. \end{aligned}$$

Solving this linear stochastic differential equation gives the result.

(iii) Given that,

$$d \log \frac{P(G \in dx \mid \mathcal{F}_t)}{P(G \in dx)} = \theta_t^{G|F}(x)\, dW_t^{\mathbb{F}} - \frac{1}{2} \|\theta_t^{G|F}(x)\|^2 dt$$

it follows that $\mathcal{D}_t \log \frac{P(G \in dx \mid \mathcal{F}_t)}{P(G \in dx)} = d\left[\log \frac{P(G \in dx \mid \mathcal{F}_t)}{P(G \in dx)}, W^{\mathbb{F}}\right]_t / dt = \theta_t^{G|F}(x)$.

(ii) Note that $dS_t/S_t = r_t dt + \sigma_t\left(\left(\theta_t^{\mathbb{F}} + \theta_t^{\mathbb{G},\mathbb{F}}\right) dt + dW_t^{\mathbb{G},\mathbb{F}}\right)$ where $dW_t^{\mathbb{G},\mathbb{F}} = dW_t^{\mathbb{F}} - \theta_t^{\mathbb{G},\mathbb{F}} dt$. As $E\left[dW_t^{\mathbb{G},\mathbb{F}} \middle| \mathcal{G}_t\right] = E\left[dW_t^{\mathbb{F}} \mid \mathcal{G}_t\right] -\theta_t^{\mathbb{G},\mathbb{F}} dt = \theta_t^{\mathbb{G},\mathbb{F}} dt - \theta_t^{\mathbb{G},\mathbb{F}} dt = 0$ and $d[W^{\mathbb{G},\mathbb{F}}]_t = d\left[W^{\mathbb{F}}\right]_t = dt$, it follows from Lévy's theorem that $\mathbb{W}^{\mathbb{G},\mathbb{F}}$ is a $\mathbb{G}$-Brownian motion. As the Doob-Meyer decomposition of the stock return under $\mathbb{G}$ is unique, it follows that $\theta_t^{\mathbb{G}} = \theta_t^{\mathbb{F}} + \theta_t^{\mathbb{G},\mathbb{F}}$. Hence, the PIPR represents the incremental price of risk.

(i) Using $\theta_t^{G|F}(x) = \mathcal{D}_t \log \frac{P(G \in dx \mid \mathcal{F}_t)}{P(G \in dx)}$ gives,

$$\int_{\mathbb{R}} \theta_t^{G|F}(x) P(G \in dx) = \int_{\mathbb{R}} \mathcal{D}_t P(G \in dx \mid \mathcal{F}_t) = \mathcal{D}_t \int_{\mathbb{R}} P(G \in dx \mid \mathcal{F}_t) = \mathcal{D}_t 1 = 0.$$

■

9
Evolutionary Behavioral Finance

Igor Evstigneev, Thorsten Hens and Klaus Reiner Schenk-Hoppé

1 The main goals and methodology

1.1 Economics and finance: global challenges

The creation and protection of financial wealth is one of the most important roles of modern societies. People will commit to working hard and saving for future generations only if they can be sure that the efforts they exert every day will be rewarded by a better standard of living. This, however, can only be achieved with a well-functioning financial market. Unfortunately, a breakdown of the financial system as in the great financial crisis of 2007 and 2008 destroys the trust in this important social arrangement. To avoid such crises we need to improve our understanding of financial markets that, so far, has been built on totally unrealistic assumptions about the behavior of people acting in them. The most fundamental and at the same time the most questionable in modern economic theory is the hypothesis of *full rationality* of economic agents who are assumed to maximize their utility functions subject to their individual constraints, or in mathematical language, solve well-defined and precisely stated *constrained optimization problems*.

This paper grew out of lecture notes prepared for presentations at the UK Mathematical Finance Workshop, King's College, London, June 2013, Trimester Program "Stochastic Dynamics in Economics and Finance" at the Hausdorff Research Institute for Mathematics, University of Bonn, June 2013, ESRC Seminar "Financial Modelling Post 2008," University of Manchester, October 2013, Technion – Israel Institute of Technology, Haifa, November 2013, and London School of Economics, December 2013. The authors are grateful to the organizers of the conferences and seminars for the invitations and to the participants for helpful discussions. Financial support from the Swiss National Science Foundation grant 149856 "Behavioral Financial Markets" (2014–2016) is gratefully acknowledged.

1.2 Evolutionary behavioral finance

The general objective of this direction of research is the development of a new interdisciplinary field, *Evolutionary Behavioral Finance (EBF)*, which combines behavioral and evolutionary approaches to the modeling of financial markets. The focus of study is on fundamental questions and problems pertaining to Finance and Financial Economics, especially those related to equilibrium asset pricing and portfolio selection. Models of market dynamics and equilibrium that are developed in the framework of EBF provide a plausible alternative to the conventional approach to asset pricing based on the hypothesis of full rationality and are aimed at practical quantitative applications.

The question of price formation in asset markets is central to Financial Economics. Among the variety of approaches addressing this question, one can observe two general and well-established theories: one deals with basic assets and the other focuses on derivative securities. Models for the pricing of derivative securities were developed in the last three or four decades, following the "Black–Scholes revolution." They were based on new ideas, led to the creation of a profound mathematical theory and became indispensable in practice. At the same time, the only general theory explaining the formation of the prices of basic assets, whether stock or equity, appears to be the Arrow-Debreu *General Equilibrium (GE)* analysis in a financial context (Radner [53]). It relies upon the Walrasian paradigm of fully rational utility maximization, going back to Leon Walras, one of the key figures in the economic thought of the 19th century. Although equilibrium models of this kind currently serve as the main framework for teaching and research on asset pricing, they do not provide tools for practical quantitative recommendations; moreover, they do not reflect a number of fundamental aspects of modern financial markets. Their crucial drawback is that they do not take into account the enormous variety of patterns of real market behavior irreducible to individual utility maximization, especially those of an evolutionary nature: growth, domination and survival.

1.3 GE theory for the 21st century

EBF develops an alternative equilibrium paradigm, which can be called *Behavioral Equilibrium*, which abandons the hypothesis of full rationality and admits that market participants may have a whole range of patterns of behavior depending on their individual psychology. Investors' strategies may involve, for example, mimicking, satisficing, and rules of thumb based on experience. They might be *interactive* – depending on the behavior of the others, and *relative* – taking into account the comparative performance of the others. Objectives of the market participants might be of an *evolutionary* nature: *survival* (especially in crisis environments), *domination* in a market segment, or fastest capital *growth*.

The evolutionary aspect is in the main focus of the models developed in the EBF. A synthesis of behavioral and evolutionary approaches makes it possible to obtain rigorous mathematical results, identifying strategies that guarantee survival or domination in a competitive market environment.

In the EBF models, the notion of a short-run price equilibrium is defined directly in terms of a strategy profile of the agents, and the process of market dynamics is viewed as a sequence of consecutively related short-run equilibria. Uncertainty on asset payoffs at each period is modeled via an exogenous discrete-time stochastic process governing the evolution of the states of the world. The states of the world are meant to capture various macroeconomic and business cycle variables that may affect investors' behavior. The traders use general, adaptive strategies (portfolio rules), distributing their current wealth between assets at every period, depending on the observed history of the game and the exogenous random factors. One of the central goals is to identify investment strategies that guarantee the long-run *survival* of any investor using them, in the sense of keeping a strictly positive, bounded away from zero, share of market wealth over the infinite time horizon, irrespective of the investment strategies employed by the other agents in the market. Remarkably, it turns out to be possible to provide a full characterization of such strategies, give explicit formulas for them and show that they are essentially, within a certain class, asymptotically unique.

This approach eliminates a number of drawbacks of the conventional theory. In particular, it does not require the assumption of *perfect foresight* (see Magill and Quinzii [44], p. 36) to establish an equilibrium, and most importantly, the knowledge of unobservable individual agents' utilities and beliefs to compute it. It is free of such "curses" of GE as *indeterminacy* of temporary equilibrium and the necessity of *coordination* of plans of market participants, which contradicts the very idea of equilibrium decentralization. It opens up new possibilities for the modeling of modern financial markets, in particular on the global level, where objectives of an evolutionary nature play a major role.

The roots of ideas underlying the EBF models lie in *Evolutionary Economics* (Alchian [1], Nelson and Winter from the 1970s to 1990s), *Behavioral Economics* (Tversky, Kahneman and Smith[1]), and *Behavioral Finance* (Shiller[2], e.g. [58]). The first mathematical models for EBF were developed during the last decade by the authors of this paper and their collaborators [4–6, 9, 23–30, 35, 36]. This research has already led to a substantial impact in the financial industry [23].

1.4 Levels of behavioral modeling

It is important to distinguish between two different methodological levels of behavioral modeling:

- *Individual level* – analyzing individual's behavior in situations involving risk.
- *Interactive level* – taking into account the dependence of an individual's actions on the actions of others and their influence on market dynamics and equilibrium.

Modern behavioral economics and finance originated from models analyzing an individual's behavior in situations involving risk. Inspired by the seminal work of Kahneman and Tversky [38], these ideas were developed in finance by Barberis and Thaler [12], Barberis et al. [10], Barberis et al. [11], Barberis and Xiong [13] and others. According to the classical approach, a decision-maker facing uncertainty maximizes expected utility. This hypothesis leads to a number of paradoxes and inconsistencies with reality (Friedman and Savage [31], Allais [3], Ellsberg [22], and Mehra and Prescott [49]). To resolve the paradoxes, the individual-level behavioral approach suggests replacing expected utilities with more general functionals of random variables, based, in particular, on the Kahneman and Tversky [38] *prospect theory* and involving *distorted probabilities* or *capacities* (non-additive measures) e.g. Denneberg [21]. In quantitative finance, studies along these lines have been pursued by Hens, Levy, Zhou, De Giorgi, Legg, Rieger, and others [18, 19, 20, 42, 61].

1.5 A synthesis of evolutionary and dynamic games

The analysis of market behavior from the angle of interaction, especially *strategic interaction,* of economic agents is a deeper and more advanced aspect of behavioral modeling. This is the primary emphasis of the research area discussed in this chapter. To build models of strategic behavior in financial markets we propose new mathematical frameworks *combining elements of stochastic dynamic games and evolutionary game theory*.

The main strategic framework of our behavioral equilibrium models is that of stochastic dynamic games (Shapley [57]). However, the emphasis on questions of survival and extinction of investment strategies in a market selection process links our work to evolutionary game theory (Weibull [60], Hofbauer and Sigmund [37]). The latter was designed initially for the modeling of biological systems and then received fruitful applications in economics. The notion of a survival portfolio rule, which is stable with respect to the market selection process, is akin to the notions of evolutionary stable strategies (ESS) introduced by Maynard Smith and Price [46] and Schaffer [55]. However, the mechanism of market selection in our models is radically distinct from the typical schemes of evolutionary game theory, in which repeated random matchings of species or agents in large populations result in their survival or extinction in the long run. Standard frameworks considered in that field deal with models based on a given static game, in terms of which the process of evolutionary dynamics is defined. Players in such models follow relatively simple predefined algorithms, which completely describe their behavior. Our model is quite different in its essence. Although the game solution concept we deal with – a *survival strategy* – is of an evolutionary nature, the notion of strategy we use is the one which is characteristic for the conventional setting of dynamic stochastic games. A strategy, in this setting, is a general rule prescribing what action to take based on the observation of all the previous play and the history of random states of the world. Players are allowed to use any rule of this kind, possess all information needed for this purpose and have a clear goal: guaranteed survival. Thus, the model at hand connects two basic paradigms of game theory: evolutionary and dynamic games [5, 6].

1.6 Unbeatable strategies

The present game-theoretic framework has the following remarkable feature. One can equivalently reformulate the solution concept of a survival strategy in terms of the wealth process of a player, rather than in terms of his market share process. A strategy guarantees survival if and only if it guarantees the fastest asymptotic growth of wealth (almost surely) of the investor using it. This can be expressed by saying that the strategy is *unbeatable* in terms of the growth rate of wealth.

Nowadays, the Nash equilibrium is the most common game solution concept. However, in the early days of game theory, the idea of an *unbeatable* (or *winning*) strategy was central to the field. At those times, solving a game meant primarily finding a winning strategy. This question was considered in the paper by Bouton [16], apparently the earliest mathematical paper in the field. Borel [15] wrote: One may propose to investigate whether it is possible to determine a method of play better than all others; i.e., one that gives the player who adopts it a superiority over every player who does not adopt it. It is commonly viewed that finding an unbeatable strategy is a problem of extreme complexity that can be solved only in some exceptional cases for some artificially designed games, such as Bouton's game "Nim." However, in our practice-motivated context, this problem does have a solution, and this is apparently one of the first, if not the first, result of this kind possessing quantitative real-world application.

The remainder of this paper is organized as follows. In Sections 2 and 3 we present the basic EBF model and the main results related to it. Section 4 focuses on a simplified version of the basic model. In the last section we discuss some open problems and directions of further research.

2 The basic model

2.1 The data of the model

In this section we present (in a somewhat simplified form) the main EBF model and key results related to it. Let $s_t \in S$ $(t = 1,2,\dots)$ be a stochastic process with values in a measurable space S. Elements in S are interpreted as *states of the world*, s_t being the state at date t. In the market under consideration, K *assets* $k = 1,\ldots,K$, are traded. At date t one unit of asset k pays *dividend* $D_{t,k}(s^t) \geq 0$ depending on the history $s^t = (s_1,\ldots,s_t)$ of states of the world by date t. It is assumed that

$$\sum_{k=1}^{K} D_{t,k}(s^t) > 0 \quad \text{for all } t, s^t,$$

$$ED_{t,k}(s^t) > 0, \quad k = 1,\ldots,K,$$

where E is the expectation with respect to the underlying probability P. Thus, at least one asset pays a strictly positive dividend at each state of the world and the expected dividends for all the assets are strictly positive.

Asset supply is exogenous: the total mass (the number of "physical units") of asset k available at date t is $V_{t,k} = V_{t,k}(s^t)$.

2.2 Investors and their portfolios

There are N *investors (traders)* $i \in \{1,\ldots,N\}$. Every investor i at each time $t = 0,1,2,\ldots$ selects a *portfolio*

$$x_t^i = (x_{t,1}^i,\ldots,x_{t,K}^i) \in \mathbb{R}_+^K,$$

where $x_{t,k}^i$ is the number of units of asset k in the portfolio x_t^i. The portfolio x_t^i for $t \geq 1$ depends, generally, on the current and previous states of the world:

$$x_t^i = x_t^i(s^t), \quad s^t = (s_1,\ldots,s_t).$$

2.3 Asset prices

We denote by $p_t \in \mathbb{R}_+^K$ the vector of market prices of the assets. For each $k = 1,\ldots,K$, the coordinate $p_{t,k}$ of $p_t = (p_{t,1},\ldots,p_{t,K})$ stands for the price of one unit of asset k at date t. The scalar product

$$\langle p_t, x_t^i \rangle := \sum_{k=1}^{K} p_{t,k} x_{t,k}^i$$

expresses in terms of the prices $p_{t,k}$ the value of the investor i's portfolio x_t^i at date t.

2.4 The state of the market

The state of the market at each date t is characterized by a set of vectors

$$(p_t, x_t^1, \ldots, x_t^N),$$

where p_t is the vector of asset prices and $x_t^1,\ldots,x_t^N$ are the portfolios of the investors.

2.5 Investors' budgets

At date $t = 0$ investors have initial endowments: amounts of cash $w_0^i > 0$ $(i = 1,2,\ldots,N)$. These initial endowments form the traders' budgets at date 0. Trader i's budget at date $t \geq 1$ is

$$B_t^i(p_t, x_{t-1}^i) := \langle D_t(s^t) + p_t, x_{t-1}^i \rangle,$$

where

$$D_t(s^t) := (D_{t,1}(s^t), \ldots, D_{t,K}(s^t)).$$

It consists of two components: the dividends $\langle D_t(s^t), x_{t-1}^i \rangle$ paid by yesterday's portfolio x_{t-1}^i and the market value $\langle p_t, x_{t-1}^i \rangle$ of the portfolio x_{t-1}^i expressed in terms of today's prices p_t.

2.6 Investment rate

A fraction α of the budget is invested into assets. We will assume that the *investment rate* $\alpha \in (0,1)$ is a fixed number, the same for all the traders. The number $1 - \alpha$ can represent, e.g., the *tax rate* or the *consumption rate*. The assumption that $1 - \alpha$ is the same for all the investors is quite natural in the former case. In the latter case, it might seem restrictive, but in the present context it is indispensable, since we focus in this work on the analysis of the comparative performance of trading strategies (portfolio rules) in the long run. Without this assumption, an analysis of this kind does not make sense: a seemingly worse performance of a portfolio rule might be simply due to a higher consumption rate of the investor.

2.7 Investment proportions

For each $t \geq 0$, each trader $i = 1,2,\ldots,N$ selects a vector of *investment proportions* $\lambda_t^i = (\lambda_{t,1}^i, \ldots, \lambda_{t,K}^i)$ according to which he/she plans to distribute the available budget between assets. Vectors λ_t^i belong to the unit simplex

$$\Delta^K := \{(a_1, \ldots, a_K) \geq 0 : a_1 + \cdots + a_K = 1\}.$$

The vectors λ_t^i represent the players' (investors') *actions* or *decisions*.

2.8 Investment strategies (portfolio rules)

How do investors select their investment proportions? To describe this we use a game-theoretic approach: decisions of players are specified by their strategies. The notion of a (pure) strategy we use is standard for stochastic dynamic games. *A strategy in a stochastic dynamic game is a rule prescribing how to act based on information about all the previous actions of the player and his rivals, as well as information about the observed random states of the world.*

A formal definition is as follows. A *strategy (portfolio rule)* Λ^i of investor i is a sequence of measurable mappings

$$\Lambda_t^i(s^t, H^{t-1}), \quad t = 0, 1, \ldots,$$

assigning to each history $s^t = (s_1, \ldots, s_t)$ of states of the world and each *history of the game*

$$H^{t-1} := \{\lambda_m^i : i = 1, \ldots, N, m = 0, \ldots, t-1\}$$

the vector of investment proportions $\lambda_t^i = \Lambda_t^i(s^t, H^{t-1})$.

Since the sets of investors' portfolios

$$x_0, x_1, \ldots, x_{t-1}, \quad x_l = (x_l^1, \ldots, x_l^N),$$

and the equilibrium prices $p_0, \ldots, p_{t-1}$ are determined by the vectors of investment proportions λ_m^i, $i = 1, \ldots, N$, $m = 0, \ldots, t-1$, the history of the game contains information about the whole *market history* $(p_0, x_0), \ldots, (p_{t-1}, x_{t-1})$.

2.9 Basic strategies

Among general portfolio rules, we will distinguish those for which Λ_t^i depends only on s^t and does not depend on the market history $(p^{t-1}, x^{t-1}, \lambda^{t-1})$. Clearly, they require substantially less information than general strategies! We will call such portfolio rules *basic*. They play an important role in the present work: the survival strategy we construct belongs to this class.

2.10 Investor *i*'s demand function

Given a vector of investment proportions $\lambda_t^i = (\lambda_{t,1}^i, \ldots, \lambda_{t,K}^i)$ of investor i, his demand function is

$$X_{t,k}^i(p_t, x_{t-1}^i) = \frac{\alpha \lambda_{t,k}^i B_t^i(p_t, x_{t-1}^i)}{p_{t,k}}.$$

where α is the investment rate.

2.11 Equilibrium and dynamics

We examine the *equilibrium market dynamics*, assuming that, in each time period, aggregate demand for each asset is equal to its supply:

$$\sum_{i=1}^{N} X_{t,k}^i(p_t, x_{t-1}^i) = V_{t,k}, \quad k = 1, \ldots, K.$$

(Recall that asset supply is exogenous and equal to $V_{t,k}$.)

Asset market dynamics can be described in terms of portfolios and prices by the equations:

$$p_{t,k}V_{t,k} = \sum_{i=1}^{N} \alpha\lambda_{t,k}^{i}\langle D_t(s^t) + p_t, x_{t-1}^{i}\rangle, \quad k = 1, \ldots, K; \tag{9.1}$$

$$x_{t,k}^{i} = \frac{\alpha\lambda_{t,k}^{i}\langle D_t(s^t) + p_t, x_{t-1}^{i}\rangle}{p_{t,k}}, \quad k = 1, \ldots, K,\ i = 1, 2, \ldots, N. \tag{9.2}$$

(All the variables with subscript t depend on s^t.) The vectors $\lambda_t^i = (\lambda_{t,k}^i)$ are determined recursively by the given strategy profile $(\Lambda^1, \ldots, \Lambda^N)$:

$$\lambda_t^i(s^t) := \Lambda_t^i(s^t, p^{t-1}, x^{t-1}, \lambda^{t-1}).$$

The pricing equation (9.1) has a unique solution $p_{t,k} \geq 0$ if $V_{t,k} \geq V_{t-1,k}$ (growth), or under a weaker assumption: $\alpha V_{t-1,k}/V_{t,k} < 1$. At date $t = 0$, the budgets involved in the above formulas are the given initial endowments $w_0^i > 0$.

2.12 Admissible strategy profiles

We will consider only *admissible* strategy profiles: those for which aggregate demand for each asset is always strictly positive. This guarantees that $p_{t,k} > 0$ (only in this case the above formula for $x_{t,k}^i$ makes sense). The focus on such strategy profiles will not lead to a loss in generality in the context of this work: at least one of the strategies we deal with always has strictly positive investment proportions, which guarantees admissibility.

2.13 Market shares of the investors

We are mainly interested in comparing the long-run performance of investment strategies described in terms of market shares of the investors. Investor *i*'s *wealth* at time t is

$$w_t^i = \langle D_t(s^t) + p_t, x_{t-1}^i\rangle$$

(dividends + portfolio value). Investor *i*'s *relative wealth,* or *i*'s *market share,* is

$$r_t^i = \frac{w_t^i}{w_t^1 + \cdots + w_t^N}.$$

The dynamics of the vectors $r_t = (r_t^1, \ldots, r_t^N)$ are described by the random dynamical system

$$r_{t+1}^i = \sum_{k=1}^{K} \left[\alpha\langle\lambda_{t+1,k}, r_{t+1}\rangle + (1-\alpha)R_{t+1,k}\right] \frac{\lambda_{t,k}^i r_t^i}{\langle\lambda_{t,k}, r_t\rangle}, \tag{9.3}$$

$i = 1,\ldots,N, t \geq 0$, where

$$R_{t,k} = R_{t,k}(s^t) := \frac{D_{t,k} V_{t-1,k}}{\sum_{m=1}^{K} V_{t-1,m}} \tag{9.4}$$

(relative dividends). Equations (9.3), following from (9.1) and (9.2), make it possible to determine $r_{t+1} = (r_{t+1}^1, \ldots, r_{t+1}^N)$ based on $r_t = (r_t^1, \ldots, r_t^N)$ and thus generate a random sequence $r_0, r_1, r_2, \ldots$ of the vectors of market shares of the investors.

2.14 Survival strategies

Given an admissible strategy profile $(\Lambda^1, \ldots, \Lambda^N)$, we say that the portfolio rule Λ^1 (or the investor 1 using it) *survives* with probability 1 if

$$\inf_{t \geq 0} r_t^1 > 0$$

almost surely (a.s.). This means that for almost all realizations of the process of states of the world (s_t), the market share of the first investor is bounded away from zero by a strictly positive random variable.

A portfolio rule Λ^1 is called a *survival strategy* if investor 1 using it survives with probability one *irrespective of what portfolio rules are used by the other investors* (as long as the strategy profile is admissible).

A central goal is to identify survival strategies. The main results obtained in this direction are outlined in the next section.

2.15 Marshallian temporary equilibrium

Some comments regarding the model are in order. The present model revives in a new context the *Marshallian* concept of temporary equilibrium. Our description of the dynamics of the asset market follows the ideas outlined (in the context of commodity markets) in the classical treatise by Alfred Marshall [45] "Principles of Economics", book V, chapter II "Temporary Equilibrium of Demand and Supply." This notion of temporary equilibrium is different from the one going back to Hicks and Lindahl (1930s–1940s), which prevailed in the GE literature during the 1970s–1990s (e.g., Grandmont and Hildenbrand [34] and Grandmont [33]). The former may be regarded as "equilibrium in actions," while the latter as "equilibrium in beliefs;" for a comparative discussion of these approaches see Schlicht [56].

In the model we deal with, the dynamics of the asset market is modeled in terms of a sequence of temporary equilibria. At each date t the investors' strategies $\lambda_{t,k}^i$, the asset dividends $D_k(s^t)$ and the portfolios x_{t-1}^i determine the asset prices $p_t = (p_t^1, \ldots, p_t^K)$, equilibrating asset demand and supply. The asset holdings $x_{t-1}^i = (x_{t-1,1}^i, \ldots, x_{t-1,K}^i)$ play the role of initial endowments available at the beginning of date t. The portfolios x_t^i selected by the agents in accordance with their demand functions are transferred to date $t+1$ and then in turn serve as initial endowments for the investors.

The dynamics of the asset market described above are similar to the dynamics of the commodity market outlined in the classical treatise by Alfred Marshall [45]. Marshall's ideas were introduced into formal economics by Samuelson [54].

2.16 Samuelson's hierarchy of equilibrium processes

As was noted by Samuelson [54], in order to study the process of market dynamics by using the Marshallian "moving equilibrium method," one needs to distinguish between at least two sets of economic variables changing with different speeds. Then the set of variables changing slower (in our case, the set of vectors of the traders' investment proportions) can be temporarily fixed, while the other (in our case, the asset prices p_t) can be assumed to rapidly reach the unique state of partial equilibrium. Samuelson [54, p. 33] writes about this approach:

> I, myself, find it convenient to visualize equilibrium processes of quite different speed, some very slow compared to others. Within each long run there is a shorter run, and within each shorter run there is a still shorter run, and so forth in an infinite regression. For analytic purposes it is often convenient to treat slow processes as data and concentrate upon the processes of interest. For example, in a short-run study of the level of investment, income, and employment, it is often convenient to assume that the stock of capital is perfectly or sensibly fixed.

As it follows from the above citation, Samuelson thinks about a hierarchy of various equilibrium processes with different speeds. In our model, it is sufficient to deal with only two levels of such a hierarchy.

2.17 Continuous vs discrete time

The above approach to the modeling of equilibrium and dynamics of financial markets requires discretization of the time parameter. The time interval under consideration has to be divided into subintervals during which the "slow" variables must be kept frozen, while the "fast" ones rapidly reach a unique state of equilibrium. In this connection, discrete-time settings in our field are most natural for modeling purposes, and attempts to realize similar ideas in continuous-time frameworks face serious conceptual and technical difficulties [51, 52].

3 The main results

Assumption 1. Assume that the total mass of each asset grows (or decreases) at the same constant rate $\gamma > \alpha$:

$$V_{t,k} = \gamma^t V_k,$$

where V_k $(k = 1,2,\dots,K)$ are the initial amounts of the assets. In the case of real assets – involving long-term investments with dividends (e.g., real estate, transport, communications, media, etc.) – the above assumption means that the economy under consideration is on a *balanced growth path*.

3.1 Relative dividends

Under the above assumption, the *relative dividends* of the assets $k = 1,\dots,K$ (see (9.4)) can be written as

$$R_{t,k} = R_{t,k}(s^t) := \frac{D_{t,k}(s^t)V_k}{\sum_{m=1}^{K}(s^t)V_m}, \quad k = 1,\dots,K, t \geq 1.$$

We denote by $R_t(s^t) = (R_{t,1}(s^t),\dots,R_{t,K}(s^t))$ the vector of the relative dividends of the assets $k = 1,2,\dots,K$.

3.2 Definition the survival strategy Λ^*

Put

$$\rho := \alpha/\gamma, \quad \rho_t := \rho^{t-1}(1-\rho)$$

and consider the portfolio rule Λ^* with the vectors of investment proportions

$$\lambda_t^*(s^t) = (\lambda_{t,1}^*(s^t),\dots,\lambda_{t,K}^*(s^t)),$$

$$\lambda_{t,k}^* = E_t \sum_{l=1}^{\infty} \rho_l R_{t+l,k}, \tag{9.5}$$

where $E_t(\,\cdot\,) = E(\,\cdot\,|s^t)$ is the conditional expectation given s^t; $E_0(\,\cdot\,)$ is the unconditional expectation $E(\cdot)$.

Assumption 2. There exists a constant $\delta > 0$ such that for all k and t we have

$$E_t R_{t+1,k}(s^{t+1}) > \delta \quad \text{(a.s.)}.$$

This assumption implies that the conditional expectation in the definition of $\lambda_{t,k}^*$, which is not less than $(1-\rho)E(R_{t+1,k}|s^t)$, is strictly positive a.s., and so we can select a version of this conditional expectation that is strictly positive for all s^t. This version will be used in the definition of the strategy Λ^*. It follows from the strict positivity of $\lambda_{t,k}^*$ that any strategy profile containing Λ^* is admissible.

A central result is as follows [5, Theorem 1]:

Theorem 1. *The portfolio rule* Λ^* *is a survival strategy.*

3.3 The meaning of Λ^*

The portfolio rule Λ^* defined by (9.5) combines three general principles in Financial Economics.

(a) Λ^* prescribes the allocation of wealth among assets in the proportions of their *fundamental values*: the expectations of the flows of the discounted future dividends.
(b) The strategy Λ^*, defined in terms of the *relative (weighted) dividends*, is analogous to the CAPM strategy involving investment in the *market portfolio*.[3]
(c) The portfolio rule Λ^* is related (and in some special cases reduces, see Section 4) to the *Kelly portfolio rule* prescribing to maximize the expected logarithm of the portfolio return.

Note that the main strength of the result obtained lies in the fact that the basic strategy Λ^*, requiring information only about the exogenous process of states of the world, survives in competition against *any, not necessarily basic*, strategies of the rivals, that might use all possible information about the market history and the previous actions of all the players.

3.4 Asymptotic uniqueness

As we noted, the portfolio rule Λ^* belongs to the class of basic portfolio rules: the investment proportions $\lambda_t^*(s^t)$ depend only on the history s^t of the process of states of the world, and do not depend on the market history. The following theorem [5, Theorem 2] shows that in this class the survival strategy $\Lambda^* = (\lambda_t^*)$ is essentially unique: any other basic survival strategy is asymptotically similar to Λ^*.

Theorem 2. *If* $\Lambda = (\lambda_t)$ *is a basic survival strategy, then*

$$\sum_{t=0}^{\infty} ||\lambda_t^* - \lambda_t||^2 < \infty \quad (a.s.).$$

Here, we denote by $||\cdot||$ the Euclidean norm in a finite-dimensional space. Theorem 2 is akin to various *turnpike* results in the theory of economic dynamics, expressing the idea that all optimal or asymptotically optimal paths of an economic system follow in the long run essentially the same route: the turnpike (Nikaido [50], McKenzie [48]). Theorem 2 is a direct analogue of Gale's turnpike theorem for "good paths" (Gale [32]); for a stochastic version of this result see Arkin and Evstigneev [7]).

3.5 The i.i.d. case

If $s_t \in S$ are *independent and identically distributed (i.i.d.)*, then the investment proportions

$$\lambda_{t,k}^* = \lambda_k^* = ER_k(s_t),$$

do not depend on t, and so Λ^* is a *fixed-mix (constant proportions) strategy*. Furthermore, λ^* is independent of the investment rate α. The most important feature of this result is that it indicates a constant proportions strategy, which survives in competition against any strategies with variable investment proportions depending on all information about the history of the game.

3.6 Global evolutionary stability of Λ*

Consider the i.i.d. case in more detail. This case is important for quantitative applications and admits a deeper analysis of the model. Let us concentrate on fixed-mix (constant proportions) strategies. In the class of such strategies, Λ^* is *globally evolutionarily stable* [26, Theorem 1]:

Theorem 3. *If among the N investors, there is a group using Λ^*, then those who use Λ^* survive, while all the others are driven out of the market (their market shares tend to zero a.s.).*

3.7 In order to survive you have to win!

One might think that the focus on survival substantially restricts the scope of the analysis: "one should care of survival only if things go wrong." It turns out, however, that the class of survival strategies coincides with the class of *unbeatable* strategies performing *in terms of wealth accumulation* in the long run not worse than any other strategies competing in the market. *Thus, in order to survive you have to win!*

To be more precise let us call a strategy Λ *unbeatable* if it has the following property. Suppose investor i uses the strategy Λ, while all the others $j \neq i$ use *any* strategies. Then the wealth process w_t^j of every investor $j \neq i$ cannot grow asymptotically faster than the wealth process w_t^i of investor i: $w_t^j \leq H w_t^i$ (a.s.) for some random constant H.

It is an easy exercise to show that *a strategy is a survival strategy if and only if it is unbeatable.*

3.8 Unbeatable strategies: a general definition [6]

Consider an abstract game of N players $i = 1, \ldots, N$ selecting strategies Λ^i in some sets L^i. Let $w^i = w^i(\Lambda^1, \ldots, \Lambda^N) \in \mathsf{W}$ be the *outcome* of the game for player i given the strategy profile $(\Lambda^1, \ldots, \Lambda^N)$. Suppose a preference relation

$$(w^j) \preceq (w^i) \ (w^i, w^j \in \mathsf{W})$$

is given, comparing relative performance of players i and j. A strategy Λ of player i is termed *unbeatable* if for any admissible strategy profile $(\Lambda^1, \Lambda^2, \ldots, \Lambda^N)$ in which $\Lambda^i = \Lambda$, we have

$$w^j(\Lambda^1, \Lambda^2, \ldots, \Lambda^N) \preceq w^i(\Lambda^1, \Lambda^2, \ldots, \Lambda^N) \quad \text{for all } j \neq i.$$

Thus, if player i uses Λ, he *cannot be outperformed by any of the rivals $j \neq i$, irrespective of what strategies they employ.*

3.9 Unbeatable strategies of capital accumulation

In our model, an outcome of the game for player i is the random wealth process $w^i = (w_t^i)$. The preference relation $\preceq$ is introduced as follows. For two sequences of positive random numbers (w_t^i) and (w_t^j), we define

$$(w_t^j) \preceq (w_t^i) \quad \text{iff} \quad w_t^j \leq H w_t^i \quad \text{(a.s.)}$$

for some random $H > 0$. The relation $(w_t^j) \preceq (w_t^i)$ means that (w_t^j) does not grow asymptotically faster than (w_t^i) (a.s.).

3.10 Unbeatable strategies and evolutionary game theory

The basic solution concepts in evolutionary game theory – *evolutionary stable strategies* (Maynard Smith and Price [46], Maynard Smith [47], Schaffer [55] – may be regarded as *conditionally* unbeatable strategies (the number of mutants is small enough, or they are identical). Unconditional versions of the standard ESS were considered by Kojima [40].

4 A version of the basic model: short-lived assets

4.1 Short-lived assets

We present a simplified version of the basic model in which assets "live" for only one period. This model is more amenable to mathematical analysis and makes it possible to develop a more complete and transparent theory. It has often served as a "proving ground" for testing new conjectures regarding the basic model. Finally, it clearly demonstrates links of the present line of studies to some adjacent fields of research, such as the classical capital growth theory with exogenous asset prices (Kelly [39], Latané [41], Thorp [59], Algoet and Cover [2], MacLean et al. [43]).

There are K assets/securities $k = 1, 2, \ldots, K$. They are issued at the beginning of each time interval $t-1, t$, yield payoffs $A_{t,k}(s^t)$ at the end of it and then expire. They are identically "re-born" at the next date t, and the cycle repeats. Asset supply at date t is $V_{t,k}(s^t) > 0$. It is assumed that

$$\sum_{k=1}^{K} A_{t,k}(s^t) > 0 \text{ for all } t \text{ and } s.$$

4.2 Investors, portfolios and prices

Investors/players $i = 1, \ldots, N$ construct *portfolios* $x_t^i = (x_{t,1}^i, \ldots, x_{t,K}^i)$ by selecting vectors of *investment proportions* $\lambda_t^i = (\lambda_{t,1}^i, \ldots, \lambda_{t,K}^i) \in \Delta^K$. The number $\lambda_{t,1}^i$ indicates the fraction of the *budget*

$$\langle A_t, x_{t-1}^i \rangle, \; A_t(s^t) := (A_{t,1}(s^t), \ldots, A_{t,K}(s^t)),$$

of investor i allocated to asset k. Note that, in contrast with the basic model, *all* the budget is used for investment. The budget at date $t = 0$ is the *initial endowment* $w_0^i > 0$.

Investors' portfolios $x_t^i = (x_{t,1}^i, \ldots, x_{t,K}^i)$ are expressed as

$$x_{t,k}^i = \frac{\lambda_{t,k}^i \langle A_t, x_{t-1}^i \rangle}{p_{t,k}},$$

and *equilibrium asset prices* $p_t = (p_{t,1}, \ldots, p_{t,K})$ are obtained from the market clearing condition (supply = demand):

$$\sum_{i=1}^{N} \frac{\lambda_{t,k}^i \langle A_t, x_{t-1}^i \rangle}{p_{t,k}} = V_k, \quad k = 1, 2, \ldots, K.$$

Thus

$$p_{t,k} = \sum_{i=1}^{N} \frac{\lambda_{t,k}^i \langle A_t, x_{t-1}^i \rangle}{V_k}.$$

4.3 Strategies and market dynamics

Vectors of investment proportions λ_t^i are selected by investors $i = 1, \ldots, N$ according to *strategies*

$$\Lambda_t^i(s^t, \lambda^{t-1}), \quad t = 1, 2, \ldots$$

depending on the history of states of the world $s^t = (s_1, \ldots, s_t)$ and the history of the game

$$\lambda^{t-1} := (\lambda_l^i), \quad i = 1, \ldots, N, \; l = 0, \ldots, t-1.$$

Basic strategies depend only on s^t, and do not depend on λ^{t-1}. A strategy profile of investors generates, as in the basic model, wealth processes of the investors

$$w_t^i = w_t^i(s^t) := \langle A_t(s^t), x_{t-1}^i(s^{t-1}) \rangle, \quad i = 1, 2, \ldots, N,$$

which in turn determine the dynamics of their market shares

$$r_t^i := w_t^i / w_t, \quad w_t := \sum_{i=1}^{N} w_t^i.$$

In the present model, the dynamics of the vectors of investors' market shares $r_t = (r_t^i, \ldots, r_t^N)$ is governed by the random dynamical system

$$r_{t+1}^i = \sum_{k=1}^{K} R_{t+1,k} \frac{\lambda_{t,k}^i r_t^i}{\langle \lambda_{t,k}, r_t \rangle}, \quad i = 1, \ldots, N, \tag{9.6}$$

which is substantially simpler than (9.3), in particular, because r_{t+1} is obtained from r_t through an explicit formula (in contrast with the implicit relation (9.3)).

It should be noted that (9.3) reduces to (9.6) when $\alpha = 0$. The intuitive meaning of this fact is clear: in the short-lived asset case one cannot reinvest. Assets live only one period and tomorrow's assets are not the same as today's.

4.4 Results

Define the *relative payoffs* by

$$R_{t,k}(s^t) := \frac{A_{t,k}(s^t)V_{t-1,k}(s^{t-1})}{\sum_{m=1}^{K}(s^t)V_{t-1,m}(s^{t-1})},$$

and put $R_t(s^t) = (R_{t,1}(s^t), \ldots, R_{t,K}(s^t))$. Consider the basic strategy $\Lambda^* = (\lambda_t^*)$ defined by

$$\lambda_t^*(s^t) := E_t R_{t+1}(s^{t+1}),$$

where $E_t(\,\cdot\,) = E(\,\cdot\,|s^t)$ is the conditional expectation given s^t. Assume

$$E \ln E_t R_{t+1,k}(s^{t+1}) > -\infty.$$

Theorems 1–3, reformulated literally for the present model, are valid (see [6, 24]).

Theorem 4. *The portfolio rule Λ^* is a survival strategy. It is asymptotically unique in the class of basic strategies. In the i.i.d. case, it is globally asymptotically stable.*

4.5 Betting your beliefs

The strategy Λ^* prescribes to invest in accordance with the proportions of the (conditionally) expected relative payoffs. This investment principle is sometimes referred to as "betting your beliefs." The same principle is in a sense valid in the basic model (see the definition of Λ^* in the previous section).

4.6 Horse race model

Consider the following toy model of an asset market (cf. Kelly [39], Blume and Easley [14]). The state space S consists of K elements: $S = \{1, 2, \ldots, K\}$, $A_k(s) = 0$ if $s \neq k$ and $A_k(s) = 1$ if $s = k$, and $V_{t,1} = V_{t,2} = \ldots = V_{t,K} = 1$. Thus, there are as many states of the world as there are assets, and one and only one asset yields unit payoff in each state of the world. Assets with this payoff structure are called *Arrow securities*.

One can think of this model as describing a sequence of horse races with independent outcomes. Only one horse k wins in each race yielding unit payoff. This event occurs with probability $\pi_k = P\{s_t = k\}$. In this example, the relative payoffs $R_k(s)$ coincide with $A_k(s)$, and the strategy $\Lambda^* = (\lambda^*)$ of "betting your beliefs" takes on the form:

$$\lambda^* = (\lambda_1^*, \ldots, \lambda_K^*), \text{ where } \lambda_k^* = ER_k(s_t) = p_k.$$

4.7 The strategy Λ^* and the Kelly portfolio rule

It is well known and easy to prove that the function

$$\Phi(\lambda) = E \ln \sum_k R_k(s_t)\lambda_k = \sum_k p_k \ln \lambda_k$$

attains its maximum over $\lambda \in \Delta^K$ at $\lambda^* = (p_1, \ldots, p_K)$. The investment strategy maximizing the expected logarithm of the portfolio return is called the *Kelly portfolio rule* (Kelly [39], Latané [41], Thorp [59], Algoet and Cover [2], MacLean et al. [43]). Thus, in the example under consideration, the strategy Λ^* coincides with the Kelly rule. It is important to emphasize that this is a specific feature of the particular case under consideration. In the general case, Λ^* is a solution to a certain game, rather than a single-player optimization problem, and a direct counterpart of the Kelly rule does not exist.

5 Problems and prospects

We list several major topics for further research.

- Developing EBF models with endogenous asset supply, short selling and leverage.
- Constructing "hybrid" models in which assets with endogenous equilibrium prices, as well as assets with exogenous prices, are traded. The role of an asset of the latter type can be played, e.g., by cash with an exogenous (random or non-random) interest rate. Some progress in the analysis of such models was made in [28].
- Developing "overlapping generations" models with a countable number of assets $k = 1, 2, \ldots$, each of which has its own life cycle starting from some moment of time σ_k and terminating at some later moment of time τ_k.
- Introducing the dependence of the dividends paid off at the end of the time period on the equilibrium prices, and consequently on the total investment in the asset expressed in terms of these prices.
- Obtaining quantitative results on the *rates* of survival and extinction of portfolio rules in the spirit of those in [9].
- Using the dynamic frameworks which are considered in EBF in more traditional settings: in models with finite time horizons and conventional solution concepts (utility maximization, Nash equilibrium).
- Introducing transaction costs and portfolio constraints into EBF models.
- Creating a *universal* version of EBF, that does not assume the knowledge of underlying probability distributions, similar to the theory of Cover's [17] *universal portfolios*.
- Conducting a systematic analysis of the notion of an unbeatable strategy in a modern game-theoretic perspective.

The above problems constitute a vast research program requiring substantial effort over a considerable time period. This program might need for its realization the development of new conceptual ideas, modeling approaches and mathematical techniques. We do not expect in the nearest perspective significant progress in all of the above directions of research, but we do expect substantial achievements in several of them, where some preliminary results have already been obtained.

Notes

[1] Kahneman and Smith: the 2002 Nobel Laureates in Economics.

[2] The 2013 Nobel Prize in Economics.

[3] However, it should be emphasized that instead of weighing assets according to their prices, in Λ^* the weights are based on fundamentals. In practice, Λ^* is an example of *fundamental indexing* (Arnott, Hsu and West [8]).

References

[1] Alchian, A.A. (1950) Uncertainty, Evolution, and Economic Theory. *Journal of Political Economy* 58, 211–221.

[2] Algoet, P.H., and T.M. Cover (1988) Asymptotic Optimality and Asymptotic Equipartition Properties of Log-optimum Investment. *Annals of Probability* 16, 876–898.

[3] Allais, M. (1953) Le comportement de l'homme rationnel devant le risque, critique des postulats et axiomes de l'ecole americaine. *Econometrica* 21, 503–546.

[4] Amir, R., Evstigneev, I.V., Hens, T., and K.R. Schenk-Hoppé (2005) Market Selection and Survival of Investment Strategies. *Journal of Mathematical Economics* 41 *(Special Issue on Evolutionary Finance)*, 105–122.

[5] Amir, R., Evstigneev, I.V., Hens, T., and L. Xu (2011) Evolutionary Finance and Dynamic Games. *Mathematics and Financial Economics* 5 *(Special Issue on Stochastic Financial Economics)*, 161–184.

[6] Amir, R., Evstigneev, I.V., and K.R. Schenk-Hoppé (2013) Asset Market Games of Survival: A Synthesis of Evolutionary and Dynamic Games. *Annals of Finance* 9, 121–144.

[7] Arkin V.I., and I.V. Evstigneev (1987) *Stochastic Models of Control and Economic Dynamics*, Academic Press.

[8] Arnott, R.D., Hsu, J.C., and J.M. West (2008) *The Fundamental Index: A Better Way to Invest*, Wiley.

[9] Bahsoun, W., Evstigneev, I.V., and L. Xu (2011) Almost Sure Nash Equilibrium Strategies in Evolutionary Models of Asset Markets. *Mathematical Methods of Operations Research* 73, 235–250.

[10] Barberis, N., Huang, M., and T. Santos (2001) Prospect Theory and Asset Prices. *Quarterly Journal of Economics* 116, 1–53.

[11] Barberis, N., Shleifer, A., and R. Vishny (1998) A Model of Investor Sentiment. *Journal of Financial Economics* 49, 307–343.

[12] Barberis, N., and R. Thaler (2003) A Survey of Behavioral Finance. In: *Handbook of the Economics of Finance* (Constantinides, G.M., Harris, M., and R. Stulz, eds), Elsevier.

[13] Barberis, N., and W. Xiong (2009) What Drives the Disposition Effect? An Analysis of a Long-Standing Preference-Based Explanation. *Journal of Finance* 64, 751–784.

[14] Blume, L., and D. Easley (1992) Evolution and Market Behavior. *Journal of Economic Theory* 58, 9–40.

[15] Borel, E. (1953) The Theory of Play and Integral Equations with Skew Symmetric Kernels. *Econometrica* 21, 97–100.

[16] Bouton, C.L. (1901–1902) Nim, A Game with a Complete Mathematical Theory. *Annals of Mathematics* (Second Series) 3, No. 1/4, 35–39.

[17] Cover, T.M. (1991) Universal Portfolios. *Mathematical Finance* 1, 1–29.

[18] De Giorgi, E., and T. Hens (2006) Making Prospect Theory Fit for Finance. *Financial Markets and Portfolio Management* 20, 339–360.

[19] De Giorgi, E., Hens, T., and M.O. Rieger (2010) Financial Market Equilibria with Cumulative Prospect Theory. *Journal of Mathematical Economics* 46 *(Mathematical Economics: Special Issue in honour of Andreu Mas-Colell, Part 1)*, 633–651

[20] De Giorgi, Legg (2012) Dynamic Portfolio Choice and Asset Pricing with Narrow Framing and Probability Weighting. *Journal of Economic Dynamics and Control* 36, 951–972.

[21] Denneberg, D. (1994) *Non-additive measure and integral*, Kluwer Academic Publishers, Dordrecht.

[22] Ellsberg D. (1961) Risk, Ambiguity and the Savage Axioms. *Quarterly Journal of Economics* 75, 643–669.

[23] Evstigneev, I.V. (2014) Mathematical Behavioral Finance: Applications in the Financial Industry, *UK Research Excellence Framework 2014*, Impact Case Study, University of Manchester.

[24] Evstigneev, I.V., Hens, T., and K.R. Schenk-Hoppé (2002) Market Selection of Financial Trading Strategies: Global Stability. *Mathematical Finance* 12, 329–339.

[25] Evstigneev, I.V., Hens, T., and K.R. Schenk-Hoppé (2006) Evolutionary Stable Stock Markets. *Economic Theory* 27, 449–468.

[26] Evstigneev, I.V., Hens, T., and K.R. Schenk-Hoppé (2008) Globally Evolutionarily Stable Portfolio Rules. *Journal of Economic Theory* 140, 197–228.

[27] Evstigneev, I.V., Hens, T., and K.R. Schenk-Hoppé (2009) Evolutionary Finance. In: *Handbook of Financial Markets: Dynamics and Evolution* (Hens, T., and K.R. Schenk-Hoppé, eds), pp. 507–566, North-Holland.

[28] Evstigneev, I.V., Hens, T., and K.R. Schenk-Hoppé (2011) Local Stability Analysis of a Stochastic Evolutionary Financial Market Model with a Risk-Free Asset. *Mathematics and Financial Economics* 5 *(Special Issue on Stochastic Financial Economics)*, 185–202.

[29] Evstigneev, I.V., Hens, T., and K.R. Schenk-Hoppé (2011) Survival and Evolutionary Stability of the Kelly Rule. In: *The Kelly Capital Growth Investment Criterion: Theory and Practice* (MacLean, L.C., Thorp, E.O., and W.T. Ziemba, eds), pp. 273–284, World Scientific.

[30] Evstigneev, I.V., Schenk-Hoppé, K.R., and W.T. Ziemba (2013) Preface to the Special Issue "Behavioral and Evolutionary Finance," *Annals of Finance* 9, 115–119.

[31] Friedman, M. and L.J. Savage (1948) The Utility Analysis of Choices Involving Risk. *Journal of Political Economy* 56, 279–304.

[32] Gale, D. (1967) On Optimal Development in a Multi-sector Economy. *Review of Economic Studies* 34, 1–18.

[33] Grandmont, J.-M., ed. (1988) *Temporary Equilibrium*, Academic Press.

[34] Grandmont, J.-M., and W. Hildenbrand (1974) Stochastic Processes of Temporary Equilibria. *Journal of Mathematical Economics* 1, 247–277.

[35] Hens, T., and K.R. Schenk-Hoppé (2005) Evolutionary Stability of Portfolio Rules in Incomplete Markets. *Journal of Mathematical Economics* 41, 43–66.

[36] Hens, T., and K.R. Schenk-Hoppé, eds, (2009) Handbook of Financial Markets: Dynamics and Evolution. Volume in the *Handbooks in Finance* series (Ziemba, W.T., ed.), North-Holland.

[37] Hofbauer, J., and K. Sigmund (1998) *Evolutionary Games and Population Dynamics*. Cambridge University Press.

[38] Kahneman, D., and A. Tversky (1979) Prospect Theory: An Analysis of Decision under Risk. *Econometrica* 47, 263–292.

[39] Kelly, J.L. (1956) A New Interpretation of Information Rate. *Bell System Technical Journal* 35, 917–926.

[40] Kojima, F. (2006) Stability and Instability of the Unbeatable Strategy in Dynamic Processes. *International Journal of Economic Theory* 2, 41–53.

[41] Latané, H. (1959) Criteria for Choice Among Risky Ventures. *Journal of Political Economy* 67, 144–155.

[42] Levy, H., De Giorgi, E., and T. Hens (2012) Two Paradigms and Nobel Prizes in Economics: A Contradiction or Coexistence? *European Financial Management* 18, 163–182.

[43] MacLean, L.C., Thorp, E.O., and W.T. Ziemba, eds (2011) *The Kelly Capital Growth Investment Criterion: Theory and Practice*. World Scientific.

[44] Magill, M., and M. Quinzii (2002) *Theory of Incomplete Markets*, Volume 1. MIT Press.

[45] Marshall, A. (1920) *Principles of Economics*. Macmillan.

[46] Maynard Smith, J., and G.R. Price (1973) The Logic of Animal Conflict. *Nature* 246, 15–18.

[47] Maynard Smith, J. (1982) *Evolution and the Theory of Games*. Cambridge University Press.

[48] McKenzie, L.W. (1986) Optimal Economic Growth, Turnpike Theorems and Comparative Dynamics. In: *Handbook of Mathematical Economics III* (Arrow, K.J., and M.D. Intrilligator, eds), pp. 1281–1355, North-Holland.

[49] Mehra, R. and E.C. Prescott (1985) The Equity Premium: A Puzzle. *Journal of Monetary Economics* 15, 145–161.

[50] Nikaido, H. (1968) *Convex Structures and Economic Theory*. Academic Press.

[51] Palczewski, J. and K.R. Schenk-Hoppé (2010) From Discrete to Continuous Time Evolutionary Finance. *Journal of Economic Dynamics and Control* 34, 913–931.

[52] Palczewski, J. and K.R. Schenk-Hoppé (2010) Market Selection of Constant Proportions Investment Strategies in Continuous Time. *Journal of Mathematical Economics* 46, 248–266.

[53] Radner, R. (1995) Equilibrium under Uncertainty. In: *Handbook of Mathematical Economics, Vol. II* (Arrow, K.J., and M.D. Intriligator, eds). Elsevier.

[54] Samuelson, P. (1947) *Foundations of Economic Analysis*. Harvard University Press.

[55] Schaffer, M.E. (1988) Evolutionarily Stable Strategies for a Finite Population and a Variable Contest Size. *Journal of Theoretical Biology* 132, 469–478.

[56] Schlicht, E. (1985) *Isolation and Aggregation in Economics*. Springer Verlag.

[57] Shapley, L.S. (1953) Stochastic Games. *Proceedings of the National Academy of Sciences* (PNAS) 39, 1095–1100.

[58] Shiller, R.J. (2015) *Irrational Exuberance* (3rd ed.). Princeton University Press.

[59] Thorp, E.O. (1971) Portfolio Choice and the Kelly Criterion. In: *Stochastic Models in Finance* (Ziemba, W.T., and R.G. Vickson, eds), pp. 599–619, Academic Press.

[60] Weibull, J.W. (1995) *Evolutionary Game Theory*. MIT Press.

[61] Zhou, X.Y. (2010) Mathematicalising Behavioural Finance. *Proceedings of the International Congress of Mathematicians*, Hyderabad, India.

10
Post-Crisis Macrofinancial Modeling: Continuous Time Approaches

Jukka Isohätälä, Nataliya Klimenko and Alistair Milne

1 Introduction

Prior to the crisis the dominant paradigm in macroeconomic modeling was the micro-founded "New-Keynesian" DSGE model (described in many textbooks including the influential exposition of Woodford (2003)). In its most basic form this combines price-stickiness with forward looking decision making by both households and firms. This provides a tractable framework for capturing the response of output and inflation to both demand and supply shocks and explaining intuitively the transmission of monetary policy (with monetary policy characterized as a choice over rules for current and future interest rates).

DSGE models have proved to be remarkably adaptable, being easily extended in many ways, most commonly by incorporating the so-called "financial accelerator," a premium on the cost of external investment finance decreasing in firm net worth (Bernanke et al. (1999)) and hence creating an extended dynamic response to shocks. DSGE models could also be fitted closely to macroeconomic data, successfully capturing macroeconomic fluctuations observed over several past decades (as demonstrated by Smets and Wouters (2005)).

Despite these successes the crisis revealed fundamental weaknesses in this DSGE paradigm. DSGE models proved incapable of explaining the protracted decline in output and investment in the industrial countries following the crisis of 2008 (or similarly persistent declines following other previous financial crises as documented by Reinhart and Rogoff (2009)). Contrary to widespread perception, DSGE models can be relatively easily extended to incorporate banks and bank balance sheets.[1] However, even with banking and other financial frictions, DSGE models, in their usual linearized form, fail to reproduce the sudden, substantial and long-lasting changes in asset prices, output or investment inherent in the periods of financial crises including that of 2008.

The objective of this chapter is to introduce an emerging literature, pursued since the financial crisis, employing non-linear continuous time specifications of economic dynamics to capture the possibility of marked and sometimes long lasting changes in financial asset prices and asset price volatility or in real economy aggregates such as output or investment.[2] Prominent contributions to this new literature include He and Krishnamurthy (2013) and Brunnermeier and Sannikov (2014a). We aim to explain the methods used in this new literature and demonstrate how they can be applied to a range of different modeling problems. This is however not a complete review of the literature on macroeconomics with financial frictions (Brunnermeier et al. (2012) provide a more extended review than we do, discussing a wider range of macroeconomic consequences of market incompleteness with extensive references to prior literature). Our aim is more limited, providing a fairly full discussion of what we perceive as some of the key contributions and describing both the economic intuition and technical solution methods that underpin their results.

This new approach to macroeconomic modeling is still very much in its infancy and the specifications employed in this generation of models are highly stylized. One way of describing this new literature is to say that it applies the tool of continuous-time modeling widely used for derivative and other asset pricing problem to a new class of macroeconomic general equilibrium problems. This though is a bit of an oversimplification – the standard financial applications of continuous-time modeling beginning with Merton (1969, 1971) and Black and Scholes (1973) all assume complete markets. By contrast, the key underlying assumption of this new literature is market incompleteness – not all risks can be costlessly traded. The reasons for this market incompleteness are, however, not typically modeled. Instead, the focus is on the implications of market incompleteness for aggregate macrodynamics and in particular the macrodynamic role of balance sheet structure (the net worth and leverage of households, companies and financial intermediaries).

Market incompleteness can also be modeled in a discrete time setting, so why employ continuous time? The reason is that specifying the dynamics of the economy in continuous time, using diffusion processes governed by stochastic differential equations or sometimes jump processes, allows for a convenient description of the fully non-linear macrodynamics. The possible realizations of the economy are characterized by a set of differential equations[3] and the solution of these equations, subject to appropriate boundary conditions, yielding both the macroeconomic outcomes (as a function of state) and the probabilities of these outcomes occurring (that is, the "ergodic" density or the probability density function of the state variable). Knowledge of the probability distribution of states then allows the analysis of the full macroeconomic dynamics. In the models reviewed in this chapter this approach is used to characterize both the impact and persistence of fundamental shocks and how this can reproduce some characteristic crisis features.

A key determining feature of the properties of this new generation of macroeconomic models is the magnitude of shocks relative to the balance sheet constraints that arise because of market incompleteness. If these shocks are relatively small, the model dynamics are dominated by the deterministic components of equations of state motion (e.g., the planned or expected saving and investment) and the diffusion of state towards these net worth or leverage constraints occurs only rarely so that model predictions are not so very different from those of conventional macroeconomic models. In this case linearized models of the kind employed in the DSGE tradition can adequately approximate the fully non-linear solution.

However, if shocks are sufficiently large so that stochastic disturbance can on occasion become much more important than the deterministic components of state motion and net worth or leverage are pushed towards constrained levels relatively frequently, then *qualitative* changes in model predictions are possible. Agents (households, firms, governments) substantially alter their behavior, not just when the constraints are actually binding but when they are close to binding and sometimes even quite far away from these constraints. They do so in order to self-insure, offsetting the absence of markets that they would like to use to protect themselves against risk. This collective attempt to avoid risk can then in turn create feedbacks at the macroeconomic level following a large disturbance. The latter induce additional volatility of asset prices encouraging even greater self-insurance and inefficient employment of real economic resources (amplification) that potentially trigger long lasting declines of real macroeconomic aggregates (persistence) such as output, employment and investment.

In these circumstances DSGE-based linearization can no longer provide an adequate description of aggregate dynamics, as this requires explicit modeling of induced volatility rather than the trend. Note though that there is no necessary and direct relationship between the magnitude of shocks and the frequency of such crisis episodes. In many of these specifications a relatively small exogenous noise may cause agents to operate with relatively small buffers of net worth, in which case even comparatively small disturbances can result in substantial departures from the predictions of linearized macroeconomic models (this is a key finding of Isohätälä et al. (2014) and seems to be what underlies the “paradox of volatility” described, for example, by Brunnermeier and Sannikov (2014a)).

This chapter contains three main sections and provides detailed discussion of six contributions to the literature. Section 2 provides a general overview of this new literature, discussing how a combination of specific economic assumptions and modeling strategy generates results which differ sharply from more established traditions of macrodynamic modeling. Section 3 reviews a number of recent applications, some journal published, others work of our own still at working paper stage. This section is itself divided into a number of subsections: 3.1 focuses on the continuous-time modeling of the dynamics of asset prices, following the approach taken by He and Krishnamurthy (2012) and also a related problem of optimal savings and consumption in general equilibrium addressed by Isohatälä et al.; 3.2 then discusses the dynamic modeling of the interaction of sectoral balance sheets with production and investment, focusing on the work of Brunnermeier and Sannikov (2014a) and the closely related partial equilibrium model of Isohätälä et al. (2014); 3.3 then discusses the further extension of these models to an explicit treatment of financial intermediation, describing current work by Klimenko et al. (2015) and Brunnermeier and Sannikov (2014d). Section 4 offers an illustrative example of the required solution methods in the context of a simple model, a simplification of Brunnermeier and Sannikov (2014a). This section is supported by a technical appendix providing a heuristic outline of solution methods. Section 5 then discusses the substantial agenda for future research opened up by this new "post-crisis" approach to macrofinancial modeling. Section 6 concludes.

2 Strengths and weaknesses of the new literature

This section provides a general overview of the new literature on continuous-time macrofinancial dynamics. Neither the economics nor the solution methods employed in this literature are in themselves especially novel. The contribution comes from combining balance sheet restrictions, in appropriately chosen contexts, with the tools of continuous-time stochastic dynamic optimization. This section therefore proceeds by outlining the economics of this new literature comparing it with an earlier substantial body of research, dating back to the late 1980s, that addresses the aggregate implications of market incompleteness. It also offers a short discussion of the technical strengths and shortcomings of this new approach.

Most of this earlier work focused on the absence of markets for insuring idiosyncratic household labor income risks, a market incompleteness that can reduce the equilibrium real interest rate (Huggett (1993); Aiyagari (1994)) and provides one potential explanation of the incompatibility of the equity market risk premium with complete market models of household consumption-savings decisions (Mankiw (1986)).[4] The particular strand of this work closest to the new macrofinancial dynamics (initiated by Krusell and Smith (1997, 1998)) considers the dynamics of capital accumulation in economies combining uninsurable idiosyncratic shocks to employment with aggregate shocks to the productivity of capital. As with the new continuous-time macrofinancial literature there are no analytical solutions, so numerical methods must be applied. A comparison of these two sections of the literature offers useful insight into their respective strengths and weaknesses.

Macrodynamic analysis with incomplete markets is only ever tractable with strong simplifying assumptions. In the presence of market incompleteness, such as limits on individual household borrowing or frictions in access of firms to capital markets, standard aggregation results no longer hold.[5] This calls into question the appropriateness of widely employed "representative agent" models. The full solution, based on the standard assumptions of complete information and model consistent expectations, requires every decision maker to track the current state and laws of motion of the entire distribution of assets and liabilities across all individual agents. There are therefore at least as many state variables as there are agents in the economy.

The new continuous-time macrofinancial literature sidesteps this challenge of aggregation, reintroducing the representative agent by assuming either that all agents of a particular type are exactly the same, with the same tastes or technology and affected simultaneously by the same shocks (within sector homogeneity); or, by assuming that all agents of a particular type can costlessly trade all financial and real assets with each other (within sector market completeness) with often at least some assets also traded between sectors.[6] These strong assumptions have allowed these models to capture qualitative changes in aggregate behavior that arise when there is a substantial probability of balance sheet constraints binding or coming close to binding, and the possibility of feedbacks that then amplifies shocks and generates persistent fluctuations in economic aggregates and asset prices. They do though illustrate one of the main points we draw from our review: this new literature is still immature with much work yet to be done to examine how well its predictions hold in more realistic settings.

The older literature on aggregate productivity shocks and uninsurable labor income deals with this aggregation problem in a quite different way, restricting attention to particular model specifications in which the solution can be reasonably accurately approximated by individual agent decision rules based on a small number of summary statistics for the entire distribution of household wealth.

The influential contribution of Krusell and Smith (1997, 1998) was to solve such a model, with two idiosyncratic employment states (employed, unemployed) and two aggregate productivity states (high in boom, low in recession), using a numerical schema which enforced model consistent capital dynamics and demonstrating that the resulting outcome exhibited "approximate aggregation" in the sense that increasing the number of summary statistics for the wealth distribution used by households in their consumption/saving decisions beyond a small manageable number did not affect model outcomes.[7] An entire branch of literature has emerged focused on the numerical accuracy of this and other alternative algorithms for solving models of this kind (for further discussion see Algan et al. (2010); Den Haan (2010)).

A weakness of the Krusell-Smith algorithm is its model dependence.[8] While it appears to work reasonably well for particular calibrations of the specific model for which it was developed, it is far from clear that it can provide a reliable approximation to the dynamics of the kind that emerge in the new continuous time macrofinancial models we review. One limitation is that it makes no allowance for the resulting dynamic changes in interest rates or other financial asset prices consequent on changes to individual agent balance sheets. Another limitation is that there is no guarantee against the algorithm converging on a “wrong” outcome, in which the particular model simulations generated at convergence contain insufficient examples of the balance sheet constraints, leading to qualitative shifts in the decisions of households or other agents that in turn substantially influence macroeconomic dynamics.[9]

Another obvious difference is that the earlier literature on macroeconomic dynamics in the presence of market incompleteness follows the dominant practice in macroeconomic modeling of assuming that time is discrete rather than continuous. The choice between discrete and continuous time is, however, less important than might at first appear. It can admittedly be a barrier to understanding.[10] But a numerical solution using a computer always eventually requires discretization. Our view is that these two assumptions (discrete vs continuous) are complementary, each with their own strengths and weaknesses. It should be possible to state any of these models using either approach, and the choice then comes down to which is more convenient for solution and communication of results.

Continuous time diffusion has some advantages. Provided that the model can be specified with a small number of heterogeneous agents, a tractable solution can be computed using ordinary or partial different equations sidestepping concerns about the existence of a “Markovian” equilibrium. Another convenience is that all paths are continuous so there is no need to be concerned about the possibility of assets or liabilities jumping beyond constrained values.[11] Solution via ordinary or partial differential equations provides an efficient way of capturing the impact across the state space of constraints on behavior at or close to boundaries. As discussed in the next section, specification in continuous time also allows the application of the convenient method of asymptotic expansion in order to capture the singularities that can emerge when financial constraints are hit. Finally, specification in continuous time with diffusion also means that decision rules can be expressed in relatively simple terms, namely, as functions of derivatives or partial derivatives of the value function (i.e., marginal values), thereby, providing useful economic intuition that is not so easily obtainable in discrete time.

Discrete time has the advantage that the solution can be computed using the well developed and widely used tools of backward recursion. The literature offers a well developed discussion of both the existence and computation of equilibria in discrete time, including for macroeconomic models with incomplete markets.[12] There are larger amounts of available software developed for solution of discrete time models. Solution methods are now well understood both when disturbances are relatively small compared to the potential constraints, so linearization can be employed, and for many non-linear models including several state variables (dynamic stochastic macroeconomic models can now be routinely solved with four or more states). Some forms of lagged response – e.g., the policy response lags resulting from delays in the release of statistical information – are more naturally specified in discrete time.

So far our comparison of these two literatures has focused on the technical challenges of aggregation and numerical solution. Comparison of these two literatures also highlights some differences in economic assumptions. One is that in older literature, for example Krusell and Smith (1998), it is individual households who are financially constrained, whereas in Brunnermeier and Sannikov (2014a) it is the representative firm that is financially constrained (as we describe below in Section 3, they are unable to borrow more than the market value of their capital). Krusell and Smith (1998) find that these underlying financial constraints make relatively little difference to aggregate dynamics, whereas Brunnermeier and Sannikov (2014a) find that the constraints substantially reduce output and investment when firm net worth (as a proportion of the market value of the economy's capital stock) falls close to zero.

But perhaps the most important advantage of the radical simplifying assumptions made in the continuous-time macrofinancial literature is the wide range of issues that can then be addressed. This will become clearer from our review of individual models in the next section that explore the impact of constraints on households, firms and intermediaries for the dynamics of asset prices, output and investment. As we discuss in Section 5, there is scope for considerable further work of this kind on the dynamic consequences of market incompleteness for a range of other aggregate economic variables, including employment, price setting, government finances and macroeconomic policy. The price paid for these advances is not insubstantial, a clear data discrepancy at the microeconomic level since not all firms or all households are able to trade amongst each other to achieve common ratios of debt to assets (i.e., the assumption that each sector can be replaced by a single representative agent is not a realistic assumption in the context of incomplete markets). This though purchases valuable new understanding of a range of macroeconomic phenomena that are attracting attention in the wake of the global financial crisis.

3 A review of some recent continuous-time macrofinancial models

In this section we review some recent continuous-time macrofinancial models. Our discussion covers three prominent papers that have attracted widespread attention together with three papers of our own. It is organized as follows. Section 3.1 discusses how continuous time models have been used to model the dynamics of assets prices, including the return on risk-free assets and the premium on risky assets. Section 3.2 reviews implications of the dynamic allocation of productive capital for financial stability. Section 3.3 discusses extensions of these models to the explicit treatment of the banking sector.

3.1 Capital constraints and asset pricing

The series of papers developed by He and Krishnamurthy (2012, 2013) (hereafter, HK(2012) and HK(2013)) explore how market incompleteness affects the risk premium on risky assets in a Lucas Jr. (1978)-type endowment economy in which cash flow yields (dividends) on risky assets follow a random walk. The key distinguishing assumption of these models is that risky assets are held only by specialist financial intermediaries subject to agency frictions similar to those modeled in Holmstrom and Tirole (1997). Incentive compatibility (i.e., avoiding the mismanagement of assets or "shirking") requires that these intermediaries must finance their investments with a minimum proportion of their own equity. When intermediary capital is scarce, this equity capital constraint binds and works as a channel of amplification of fundamental shocks to intermediary assets and net worth, increasing the volatility of returns and the risk premium earned from investment in the risky asset.

Here we focus on the model developed in HK(2012) (the other model is similar). HK(2012) model an economy in which there is a single risky, non-tradable asset of *a fixed size* and the market price P_t that reflects the expected discounted value of dividend streams.[13] The asset generates a stochastic flow of dividends D_t per unit of time, that evolves as a Geometric Brownian motion with a constant drift and volatility σ. There is also a risk-free asset (bonds) in zero net supply and interest rate r_t, i.e., there is the possibility of lending between the households and specialists. The risky asset's risk premium is then given by

$$\pi_{R,t} = \mathbb{E}\Big[\frac{D_t\,\mathrm{d}t + \mathrm{d}P_t}{P_t}\Big]/\mathrm{d}t - r_t.$$

There are two classes of investor: *specialists* managing financial intermediaries that play the role of investment vehicles and *households* who delegate investment decisions to specialists, as they have no direct access to investment technologies (i.e., there is market segmentation). In this and all following models we review in this section, all agents belonging to a particular group are identical. Such a simplification is key for obtaining tractable solutions, as it allows working with a representative agent making the optimal decisions based on observations of her own level of wealth and (typically) a unique aggregate state.

Both specialists' and households' wealth is invested in intermediaries. The optimal contract between households and specialists determines $\beta_t \in [0,1]$ – the specialist's share of investment in the risky asset and hence, after allowing for a fee $K_t \, \mathrm{d}t$ that may be paid to specialists for managing entrusted funds, their claim on dividend income. Specialists choose the total volume of investment in the risky asset, $\mathcal{E}_t$, and make a working/shirking decision unobservable to households. As in Holmstrom and Tirole (1997), shirking reduces the cash-flow from risky assets by $X_t \, \mathrm{d}t$ but enables specialists to collect private benefits $B_t \, \mathrm{d}t$ which are assumed to be proportional to the reduction in the asset cash-flow, caused by shirking:

$$B_t \, \mathrm{d}t = \frac{1}{1+m} X_t \, \mathrm{d}t,$$

where the inverse of m captures the magnitude of agency frictions.[14]

The incentive contract preventing shirking places restrictions on outside equity financing. Namely, the households' equity stake must be limited to a fraction of the total risky investment that depends on the magnitude of agency frictions, which leads to the following equity capital constraint:

$$\mathcal{E}_t^h \leq m\mathcal{E}_t. \tag{10.1}$$

Put differently, to abstain from shirking, specialists must maintain some "skin in the game," whose proportion is increasing with the magnitude of agency frictions. In terms of the sharing rule, the above constrain implies that

$$\beta_t^* \geq \frac{1}{1+m}, \quad \text{for} \quad K_t \geq 0,$$

with equality when $K_t > 0$.

To obtain a closed form solution He and Krishnamurthy (2012) assume that both specialists and households have log-preferences over instantaneous consumption. With this assumption, the value function of any representative agent is additively separable and can be written in the following form:

$$\frac{1}{\rho^i} \log(W_t^i) + Y_t^i,$$

where ρ^i is the discount rate of the agent $i = \{s,h\}$ (specialist and household, respectively), W_t^i is the wealth of the agent i and Y_t^i is the function of the aggregate wealth and dividends, which are two state variables in this setting.

Due to the above property of the value function, the portfolio and consumption choices of agents are almost trivial. In particular, agents continuously consume an amount proportional to their net worth, where the consumption rates are given by their respective discount factors, i.e., $c_t^i = \rho^i W_t^i$ and the optimal exposure to the risky asset is given by the mean-variance portfolio choice, yet, with a slight twist for households for whom the effective asset risk premium is reduced by the scaled intermediation fees $k_t = K_t/\mathcal{E}_t^h$.

Finding the unique equilibrium of this model requires solving for three processes – risky asset price P_t, riskless interest rate r_t and scaled intermediation fees k_t – compatible with the individual maximization and market clearing conditions. P_t, r_t and k_t are the functions of the unique state variable – the aggregate specialists' wealth scaled by aggregate dividends, $w_t \equiv W_t^s/D_t$.

Depending on the level of the scaled specialists' net worth, at each moment of time the economy can find itself in one of two regimes: if scaled specialist net worth w_t exceeds a critical threshold $w^c \equiv 1/(\rho^h m + \rho^s)$ then the solution is in an unconstrained regime in which the incentive constraint (10.1) is slack; otherwise the solution is in a constrained regime in which the incentive constraint (10.1) is binding.

In the unconstrained regime where the wealth of the specialist financial intermediaries is relatively high, the risk premium is constant and households pay zero fees for intermediation. There is no borrowing or lending (with the implicit "risk-free" rate of interest r_t, a wealth dependent weighted average of the discount rates of households and specialists, that declines as w_t increases). Holdings in the risky asset (β_t) are proportional to agent wealth. The price volatility of the risky asset is constant and is equal to the volatility of the dividend cash-flow, i.e., $\sigma_{R,t} = \sigma$.[15]

In the constrained region, in which the wealth share of the specialist financial intermediary are relatively low (below w^c), the equity constraint binds. Their relatively low level of wealth means that the specialists must borrow from households in order to maintain their required share of holdings of the risky asset. The volatility of the risky asset (endogenous volatility),

$$\sigma_{R,t} = \sigma \left[\frac{(1+m)\rho^h}{(\rho^h m + \rho^s)(1 + (\rho^h - \rho^s) w_t)} \right] > \sigma,$$

then drives the level of both the risk premium and of intermediation fees in the constrained regime. These are both always higher than in the unconstrained regime, but decreasing with the scaled specialists' wealth until the threshold between the two regimes is reached. The risk-free interest rate (at least for the chosen parameterizations) also exhibits a different pattern than in the unconstrained region: namely, it becomes an increasing function of specialist wealth; i.e., in the constrained regime, the lower the specialist wealth, the higher the valuation placed on risk-free assets.

The HK(2012) model predicts that intermediaries only borrow in the constrained regime, otherwise intermediaries are unleveraged. In order to generate leverage in the unconstrained regime and so better match the data, HK(2013) amend their earlier model by introducing household labor income uncertainty and an exogenous demand by households for holding a minimum proportion of wealth in the form of risk-free lending to specialists.[16] The solution is now numerical, not closed-form. Parameters are chosen so that, absent of any constraints, the risk-tolerant households hold all their wealth in the form of risky assets and as a result the equity constraint on specialists binds approximately 50% of the time. With this set-up the model does a fairly good job of reproducing the dynamics of risk-premia during financial crises, with a "half-life" (an expected decline of the risk-premia relative to unconstrained levels of 50%) of about eight months.

Further insight into the impact of leverage constraints on the pricing of *risk-free* assets is provided by Isohätälä et al. (2015) (hereafter, IKMR(2015)). They consider the interaction of two household sectors receiving an endowment income subject to offsetting shocks: a positive shock to income and an equal and opposite negative shock in the other. There is a single consumption good. Cumulative income is a diffusion process with infinite local variation (the standard deviation of income over a period t to $t+\Delta t$ is proportional to $\sqrt{\Delta t}$ while expected income is proportional to Δt). While there is no insurance contract that protects against this income uncertainty (the assumed market incompleteness), households can still smooth consumption by borrowing and lending from each other, subject to a constraint of some maximum level of borrowing. Both households seek to maximize a standard objective, the discounted expected utility with instantaneous "CRRA" utility, i.e., constant relative risk aversion and intertemporal elasticity of substitution. One household is relatively impatient discounting consumption more than the other. The underlying microeconomics are not further developed although the constraint on borrowing might represent the possibility of repudiating debt and instead obtaining some alternative subsistence income.

These strong modeling assumptions yield a simple and intuitive outcome with buffer stock saving very similar to that predicted by standard microeconomic models of household precautionary saving. Both household consumption (c) and expected saving, i.e., expected endowment and financial income net of consumption ($a + r(w)w - c(w)$), are monotonic functions of wealth w, with consumption increasing and savings decreasing with w. Here wealth w is simply the net claims of impatient households on households in the other patient sector, so $-w$ is a measure of impatient household leverage (w is almost always negative). Expected saving by the impatient household sector is positive whenever leverage is above a target level (buffer stock saving). The novel macrofinancial feature of the model is that the real interest rate $r = r(w)$ adjusts to ensure goods market clearing: i.e., total consumption by the two sectors equals their total endowment, with potentially large but relatively short lived declines of real interest rates whenever income shocks increase the leverage of impatient households close to their maximum levels of borrowing. This is thus a setting in which a financial problem (overleverage) is corrected in large part through adjustment of market prices (a temporary period of low real interest rates supporting deleveraging towards a long term desired level of borrowing) rather than through reduction of consumption.

There are sharp contrasts between the investigations of IKMR(2015) and those of HK(2012) and HK(2013), but also striking similarities. Differences include: the specification of uncertainty (in HK(2012) this is a diffusion process for the risky asset's productivity, while in IKMR(2015) this is a diffusion process for cumulative endowment); the distinction between sectors (in HK(2012) this distinction is between specialist asset managers and outside investors, while in IKMR(2015) this distinction is between impatient borrowing households and patient lending households); the focus of the analysis (in HK(2012) this is the pricing of the risky asset while in IKMR(2015) it is the pricing of risk-free instantaneous borrowing); and in the treatment of household optimization (in HK(2013) the OLG setting abstracts from all issues intertemporal cash management while in IKMR(2015) both agents address a fully intertemporal optimization).

The key similarity is that in both settings asset prices adjust so as to restore balance sheets fairly quickly towards long run expected values. Periods of distress are relatively short lived. Following initial disturbances, after a few months wealth shares gravitate back towards the steady state distribution (the "ergodic density" across wealth). In particular, in all these settings risk-free interest rates decline dramatically during periods of extreme financial stress and this assists the process of deleveraging (see HK(2013) figure 3 and IKMR(2015) figure 6).

3.2 Models of output and investment *without* an explicit banking sector

In this section we describe the model of Brunnermeier and Sannikov (2014a) (hereafter, BS(2014-1)) that focuses on the role of net worth in the allocation of productive capital in the economy and its implications for the dynamics of output and investment. In BS(2014-1), capital is traded between more productive, risk-neutral, impatient *experts* and less productive, risk-averse, more patient *households*.[17] The productivity of capital follows a diffusion process, as in the complete market setting of Lucas Jr. (1978) and employed by HK(2012) and HK(2013). Also as in HK(2012) and HK(2013), the state of the economy is described by the single state variable, the ratio of expert net worth to household net worth.

As well as sharing in the risky investment opportunity, households may invest in risk-free debt issued by experts. Debt contracts are short term, and experts continuously adjust their level of debt in order to balance a desire to consume early (impatience) against the potential costs of incomplete insurance against productivity shocks.[18] While the BS(2014-1) model features no explicit leverage or capital constraint, a constraint emerges implicitly because reductions in the market value of capital limit the ability of firms to borrow. The absence of a market for insuring against fluctuations in the productivity of capital and hence net worth mean that, in effect, debt is subject to a collateral constraint, not unlike that featured in Kiyotaki and Moore (1997).

This implicit need for collateralization is consistent with a standard paradigm of financial intermediation literature considering financial intermediaries (particularly, banks) as the providers of safe and liquid investment opportunity (demand deposits), given that some economic agents may have strong preferences for this kind of investment (see, e.g., Diamond and Dybvig (1983), DeAngelo and Stulz (2013) for the arguments along this line). In the environment in which financial intermediaries act as the liquidity providers, while facing financial frictions, this feature creates a role for intermediaries' net worth as a loss-absorbing buffer that is needed to guarantee the safety of debt issued to households.

The productivity of capital in the BS(2014-1) economy fluctuates over time according to a diffusion process with standard deviation σ. This in turn alters both expert net worth and the share of expert net worth (a positive shock to productivity of capital increases the net worth of both experts and households; as long as experts are leveraged then this also increases the share of expert net worth). They assume in addition that new physical capital can be built via an investment technology with adjustment costs.[19] The main friction in this economy refers to the fact that experts do not have "deep pockets" and cannot raise outside equity (this, in fact, can be interpreted as the extreme form of the agency problem present in HK(2012)). As a result, a decline in net worth caused by negative productivity shocks increases the effective risk aversion of experts. This induces them to "self-insure" by shrinking the scale of operation (simultaneously, reducing the volume of debt) and selling capital to less productive households, which ultimately leads to the reductions in output. Moreover, sales of capital by experts depress the asset price, which, in turn, feeds back into the dynamics of net worth, thereby amplifying the impact of the adverse productivity shock.[20] We illustrate the detailed modeling of this mechanism in Section 4 by using a simplified version of BS(2014-1)'s model.

In equilibrium, the dynamics of capital prices, capital and experts' net worth, as well as the optimal consumption and investment decisions of agents (and their respective holdings of capital), can be characterized as the functions of a single state variable – the *experts' share in the total net worth*. Expression in terms of a single state variable is possible due to the linearity of the agents' value functions in individual agent's net worth (scale-invariance property). The optimal consumption decisions of experts (who face the non-negative consumption constraint) are determined by the marginal value of their net worth, which is a decreasing function of the state.

In the baseline model explored by BS(2014-1) the optimal consumption pattern is similar to the optimal payout policies emerging in many (partial equilibrium) corporate finance models: as long as the value of the state is relatively low and thus an expert's net worth is highly valuable, it is optimal to retain earnings; however, once the marginal value of the state falls to one, experts consume all positive profits so as to maintain the state at the level associated to the unit marginal value.[21] Such a "barrier-type" consumption strategy determines the upper bound of the state. The fluctuations of the state between zero and the consumption boundary drive the effective risk aversion of experts and thus the equilibrium allocation of capital in the economy: as long as the share of experts' net worth is relatively high, all capital is concentrated in the experts' hands; however, below a certain critical level, the fraction of capital held by experts is always lower than one and is an increasing function of the state.

An important effect captured by BS(2014-1)'s model is extended persistence of the aggregate shocks, a consequence of the response of experts to the incomplete opportunities for insurance against productivity risks. As the share of expert net worth declines and an increasing proportion of capital is sold to and managed by households, it becomes relatively difficult for experts to rebuild net worth via retained earnings. This means that for at least some parameter combinations the economy may spend quite a lot of time in recession states with low asset prices and a large fraction of capital concentrated in the hands of less productive agents. This property manifests itself via the ergodic density of the state being spiked in the neighborhood of its lower boundary.[22]

A point that is not entirely clear in BS(2014-1) is the respective importance of the "self-insurance" effect and amplification effect generated by the endogenous volatility of the price of capital in generating these protracted dynamics. Certainly it is possible to get similarly protracted dynamics without endogenous price volatility. This point is illustrated by the closely related partial equilibrium model of Isohätälä et al. (2014) (hereafter, IMR(2014)). In this paper identical impatient firms manage a risky asset and the diffusion process affects aggregate accumulated cash flow rather than the productivity of capital. Moreover, in order to reduce risk exposure, capital is rented by experts to patient households rather than sold. Preferences are the same as in the baseline model of BS(2014-1), i.e., both experts and households have linear preferences but experts are subject to a "non-negativity" constraint on consumption, i.e., in effect a prohibition on issue of new equity capital. Unlike BS(2014-1), this model also parameterizes the deadweight costs of equity issuance. The merit of this model specification is its relative simplicity and tractability, as there is no need to take any account of the complications of asset pricing or optimal portfolio allocation.

The optimal risk exposure chosen by a representative firm in IMR(2014) depends on its leverage and is implemented via the optimal rental decisions:

at each moment of time, firms may unload some risk by leasing a fraction of capital to less productive households in return for a fee (assumed equal to the productivity of capital in the hands of households).

The IMR(2014) economy exhibits a very similar behavior to the one that emerges in the BS(2014-1) setting, albeit without price volatility: under the combination of relatively high uncertainty and large financing frictions (i.e., high recapitalization costs) the economy spends a lot of time in the recession states characterized by low experts' net worth and a large fraction of capital concentrated in the hands of less productive households.

3.3 Models of output and investment *with* an explicit banking sector

In this section we consider two continuous time macrofinancial models with a more explicit treatment of the banking sector. The first is that of Klimenko et al. (2015) (hereafter, KPR(2015)) which distinguishes the banking sector from the productive sector in order to address the role of bank capital in the fluctuations of credit and output. This model captures a complementary channel for output distortions that works via the adjustments of credit volumes in the economy.[23] The second is the more ambitious modeling of Brunnermeier and Sannikov (2014d) (hereafter, BS(2014-2)) who develop a monetary analysis in which net worth limits the ability of banks to create "inside-money" and thus affects both the real economy and the nominal price level. While the dynamics of risk-premia and of output and investment generated by these models are similar to those reviewed earlier in this section, the explicit treatment of banking allows a much fuller discussion of policy instruments, including bank capital regulation, as well as monetary and fiscal policy. We should emphasize that work on both these models is ongoing – when eventually published in peer-review journals they could have evolved substantially from the versions we discuss here. Still we think these two models are worth highlighting as examples of where the continuous-time macrofinancial literature may be heading in the future.

KPR(2015) study the impact of bank capital on the cost of credit in the economy where the firms' projects are financed exclusively via bank loans. The model shares some similarities with HK(2012) and BS(2014-1). Again there are two classes of agents, in this case relatively impatient banks and relatively patient households. Banks are risk-neutral and by implication (since they are maximizing expected utility) have an infinite intertemporal elasticity of substitution. Households also have a infinite intertemporal elastisticy of substitution, with a time discount rate of r, and are willing to provide unlimited deposits at an interest rate r but only as long as there is no risk of any loss on deposits.

The economy is subject to aggregate shocks, which affect the firms' default probability (and cannot be diversified) and ultimately the banks' profits. Cumulative profits (retained earnings) are described by a diffusion process with drift and diffusion proportional to the volume of bank lending. The firms' demand for credit is an exogenous decreasing function of the nominal loan rate R_t, where the latter is determined at equilibrium as a function of aggregate bank capitalization E_t. Banks continuously adjust the volume of lending, as well as the volume of deposits they collect.[24] However, their capacity to adjust book equity (net worth) is limited, because banks face a proportional deadweight cost γ, when raising new capital (this parameterization is similar to that employed by IMF(2015)).

A convenient property of the model is a linearity of the value function of an individual bank in the level of its book equity. Banks in KPR(2015) economy behave competitively in both loan and deposit markets and make the same decisions. As a result, all banks' decisions (lending, recapitalization and dividend payouts) are driven by the market-to-book value of their equity, which is the same for all banks and a function of aggregate bank capitalization.[25]

Aggregate lending, recapitalization and dividends are then functions of the level of aggregate bank capital that follows a Markov diffusion process reflected at two boundaries: banks are paying dividends at the upper boundary and recapitalize as soon as the book equity is depleted. In other words, to reduce the frequency of costly recapitalizations, banks maintain equity buffers, whose target size is optimally chosen so as to maximize shareholder value. As a consequence of the risk-neutrality of banks, dividend behavior is of the same "barrier control" form as in the baseline model of BS(2014-1) and in IMR(2014) with payments only when bank equity climbs to an upper level E_{max}. There is also recapitalization at a lower barrier E_{min}, which turns out to be zero in the competitive equilibrium.

The value function represents the expected value of the bank shareholders' claim and can be expressed as the product of the book equity of the individual bank times the market-to-book value of the banking sector. This decomposition of the value function helps understand the source of a positive lending premium (the margin between loan and deposit rates $R_t - r > 0$ with equality at the upper dividend paying boundary) emerging from this model: any negative shock to bank earnings not only depletes book equity (directly reducing lending capacity) but is further amplified via a decline in the market-to-book value. As bank equity declines, bank shareholders become effectively more and more risk-averse (even though their preferences are risk-neutral) and demand a strictly positive premium in order to lend to the real sector. This lending premium (as well as the loan rate itself) is a decreasing function of aggregate bank capitalization. It is also a non-linear function (similar non-linearity emerges also in BS(2014a) and IMR(2014)). When aggregate bank capitalization is relatively high, close to the upper dividend paying boundary, then changes in bank capital have a relatively small impact on the lending premium. When aggregate bank capital is relatively low, close to the lower recapitalization boundary, then changes in bank capital have a relatively large impact on the lending premium.

Via this lending premium channel, the reductions in aggregate bank capitalization ultimately translate into a higher cost of credit, a reduction of the firms' demand for bank loans, and thus a decline of output.

The explicit dynamics of the loan rate that emerges in KPR(2015) model (with the further assumption that the deposit rate $r = 0$, the drift and volatility of the loan rate can be obtained in closed form) allows for a tractable analysis of the long run behavior of the economy using the loan rate as the state variable. As in the BS(2014-1) or IMR(2014) models discussed above, the dynamics can be described by the ergodic density function of the state. The analysis of the ergodic density patterns shows that the economy spends a lot of time in states with lower endogenous volatility and, under strong financing frictions can get trapped in states with low bank capitalization, a high loan rate (low lending) and thus low output.

One of the natural applications of KPR(2015) is the analysis of the impact of capital regulation on financial stability and lending. The analysis shows that increasing the minimum capital ratio drives up the loan rate and induces a substantial decline in lending in the *short run*. However, in the *long run*, this negative effect is mitigated due to enhanced financial stability, as banks spend more time in the states with higher capitalization and thus relatively low loan rates.

BS(2014-2) also develop a model in which the experts are financial intermediaries or banks. The basic assumptions are that banks have a superior monitoring technology than households (in this respect their setup is similar to that of Diamond (1984) and KPR(2015)).

Their goal is however much more ambitious than that of KPR(2015). The bank share in aggregate net worth (the usual state variable) determines the extent to which they can issue short term liabilities (inside or "i" money) and hence drive aggregate macroeconomic dynamics, both real economy output and investment and nominal pricing.

Like BS(2014-1), BS(2014-2)'s model considers two classes of agent (households and experts) but now with the same rate of time preference. The experts are now financial intermediaries, distinguished because the monitoring technology of intermediaries allows them to achieve superior performance from investment in a subset set of available technologies. Banks also benefit from diversification because they can invest in many technologies. Households in contrast can invest only in a single technology (at any point in time). The inability of households to diversify idiosyncratic risk again creates a demand for holding monetary deposits. In this model, such deposits are risky – because of the risk of changes in the nominal price level – but still carry a lower risk than any other technology in the economy.

As in BS(2014-1), experts' net worth serves as a loss-absorbing buffer. Again this because markets are incomplete and experts cannot fully insure against fluctuations in the productivity of capital. Their net worth then affects the level of "inside money" (i.e., bank deposits). This value is determined by a simple equilibrium mechanism: when negative shocks deplete the experts' net worth, in order to reduce exposure to further shocks, they shrink their balance sheet by selling capital (loans to end-borrowers) to households. Due to the balance sheet adjustment, this automatically leads to the reduction of their deposit taking capacity, i.e., the supply of inside money shrinks. However, the households' demand for deposits (money) remains almost unchanged, and hence the "price" of money in terms of goods (p) must rise at the same time. Thus a contraction of intermediary net worth both reduces the price of capital in terms of goods (q) and increases the price of money. A rise in the nominal price of money is a fall in the price of goods, so this becomes a model of *disinflation* (assuming that monetary policy i.e., the supply of outside money, remains fixed).

The BS(2014-2) model is a promising framework for a tractable analysis of macroprudential policies and both orthodox and unorthodox monetary policy. It is though difficult to relate their model to the widely accepted "new-Keynesian" treatment of monetary policy widely employed in DSGE modeling. In the "new-Keynesian" world money stocks, indeed all balance sheets, are essentially irrelevant, the main market friction is sluggishness of price adjustment usually determined in the optimization setting of Calvo (Calvo (1998)) by assuming a fraction of price-setters in imperfectly competitive final goods markets can readjust prices at any point in time (without this feature DSGE models would exhibit price-neutrality, nominal pricing and monetary policy would then be entirely irrelevant to the real economy). In conventional DSGE stocks of money (as opposed to monetary policy) play no role at all.

There are of course many macroeconomic models in which the stock of money does play an essential role. These include many models in which money is required as a means of payment, either using the relatively ad-hoc mechanism of a "cash-in-advance" constraint (Lucas and Stokey (1987)) and also search models in which money provides a solution to the problem of exchange between anonymous parties who have no mechanism to commit to contractual agreements (for example, the relatively tractable model of Lagos and Wright (2005)).

The role played by money in BS(2014-2) is not a means of payment but a store of value. In this respect its role is comparable to that in the many overlapping generation models of money originating with Samuelson (1958). The simplest example is the two period overlapping generations endowment economy with a single non-storable good. Without money younger generations are unable to lend to or borrow from the current older generation at period t in order to consume less or more than their period t endowment, the problem being that the older generation are no longer around to receive or make repayment in the following period $t+1$. The equilibrium is autarky, with each generation consuming its own current endowments. With standard assumptions about preferences there is though an alternative welfare improving equilibrium (at least one) with "money". For example young generations at t anticipating a large decline in their future endowment may save for old age by acquiring money. In the subsequent time period $t+1$ (when they themselves are old) they spend this money, acquiring goods from the new younger generation. Money serves as a store of value and allows exchange to take place because of the belief that it will have an exchange value for goods in each subsequent period. Such an equilibrium exists provided that there is no terminal time at which a new generation is no longer born and money has no value.

The demand for money as a store of value in BS(2014-2) is different from that in these overlapping generation models, arising because of the risk diversification available to households from holding money. Still, as in the substantial literature on overlapping generations models with money, this basic model in which money serves as a store of value can be extended to investigate several issues in monetary policy. Government can alter the equilibrium outcome by issuing "outside money" as an alternative store of value, entirely equivalent from the perspective of households to inside money issued by intermediaries. Government can also offer interest on this outside money and issue long term bonds. Overlapping generations models of money have been used to explore many monetary issues, including the distributional and efficiency impact of different monetary policy rules (for example providing support to the Friedman rule that dynamic efficiency requires that the supply of outside money should contract, and its value increase, at a rate equal to the equilibrium rate of interest).

The major difference and the key contribution of BS(2014-2) is that their setting incorporates business cycle fluctuations; therefore, they are able to consider the role of these various monetary policies not just in steady state, but also as a tool for countering macroeconomic fluctuations through the redistributional effect of altering the distribution of net worth between creditors (in their case households) and debtors (in their case financial intermediaries). Policies which redistribute wealth from debtors to creditors following large shocks can help limit the occurrence and duration of extended downturns (deflation) in which output decreases, the price of money p is high and the price of capital q and hence investment is low.

It is clear that there is considerable scope for further research, investigating the robustness of these BS(2014-2) findings in a range of other settings. It is possible that similar results could be obtained using other models of "inside money."[26] The question of how to integrate market incompleteness and balance sheet constraints with conventional models of monetary policy remains a central issue for future research and continuous time macrofinancial models, building further on the work of BS(2014-2), may yet provide considerable further insight.

4 An illustrative example: output in general equilibrium

The purpose of this section is to present a simple and tractable example of a continuous-time macrofinancial model, in order to illustrate both methods of solution and some of the insight that can be obtained from this kind of model. The model we present here is essentially that of BS(2014-1), but slightly simplified in that there is no investment. The solution method we apply to solve this model differs from the one employed in the original BS(2014-1) model, but leads to the same results.

We develop this example with three objectives in mind: first, it shows how financing constraints mathematically appear in continuous time general equilibrium models; second, it gives a quick recipe for numerically solving such models; finally, it provides a concrete illustration of how such a model can, at least under some parametrization, explain persistence of fundamental shocks reflected by a protracted reduction of output. The appendix to this chapter provides a short heuristic summary of the mathematical solution methods used in this literature, and further technical references containing a more rigorous presentation of these methods.

4.1 Model

In this illustrative example we consider a hypothetical economy that consists of two types of agents: experts and households (we will use an overbar to denote state variables and parameters corresponding to households). A representative expert (household) is characterized by two state variables: cash c ($\bar{c}$) and capital k ($\bar{k}$). Cash holdings earn interest at a constant exogenous rate r, while capital gives production yields at rates a and $\bar{a}$. Negative cash holdings are interpreted as debt. Agents consume their wealth at rates κ and $\bar{\kappa}$ that are to be determined by maximizing appropriate objective functions. Experts and households are identical, except for the following three differences: (*i*) households are less productive, $\bar{a} < a$, (*ii*) households are more patient than experts, which is captured by the difference in their respective discount rates $\bar{\rho} \equiv r < \rho$, and (*iii*) household consumption is not constrained, whereas an expert must have a non-negative consumption, *i.e.*, $\kappa \geq 0$.

Capital can be freely traded between experts and households at a stochastically varying price q_t. Capital does not depreciate, but is subject to productivity shocks with an amplitude σ per unit capital and square root unit time. At equilibrium, the market for capital and debt must clear.

Under the above assumptions, the expert cash and capital follow the stochastic differential equations (here for experts only, analogous equations hold for households)

$$\mathrm{d}c_t = (ak_t + rc_t - q_t\tau_t k_t - \kappa_t)\,\mathrm{d}t, \tag{10.2a}$$

$$\mathrm{d}k_t = \tau_t k_t\,\mathrm{d}t + \sigma k_t\,\mathrm{d}z_t, \tag{10.2b}$$

where τ_t is the rate at which the agent trades capital (positive τ buys, negative sells) and $\mathrm{d}z_t$ captures the aggregate productivity shocks. The capital price is supposed to be stochastic and follows the equation

$$\mathrm{d}q_t = \mu_t^q q_t\,\mathrm{d}t + \sigma_t^q q_t\,\mathrm{d}z_t, \tag{10.3}$$

with initial data q_0 and where the drift and diffusion functions μ_t^q and σ_t^q are some functions of time to be determined in equilibrium.

Since the capital trade is unconstrained, the agents are free to allocate whatever proportion of their net worth, $n_t = c_t + q_t k_t$, between the risk-free asset and capital. Let $\varphi_t = q_t k_t / n_t$ denote the proportion of an agent's net worth invested in capital (note that $c_t = (1 - \varphi_t)k_t$). Applying the Itô's Lemma to n_t [Technical Appendix A.1, Eq. (A.35)], we get

$$\mathrm{d}n_t = \left[r + \left(\frac{a}{q_t} + \mu_t^q + \sigma\sigma_t^q - r \right)\varphi_t - \lambda_t \right] n_t\,\mathrm{d}t + \left(\sigma + \sigma_t^q\right)\varphi_t n_t\,\mathrm{d}z_t. \tag{10.4}$$

Note that, for convenience, we have also re-written consumption as $\kappa_t = \lambda_t n_t$. The structure of the above equation is essentially the same as in classical Merton's portfolio problem (Merton, 1969): The agent makes the allocation choice φ between the risky (capital) and risk-free (cash) assets with the goal of maximizing the value of pay-off from a (self-financing) portfolio. The main difference pertains to the fact that the price of capital, q, does not follow a geometric Brownian motion, as coefficients μ^q and σ^q (that will be endogenously determined below) are not constant.

Following Brunnermeier and Sannikov (2014a), we hypothesize that the aggregate state of the economy is given by a one-dimensional diffusion process which we here call x, and posit the equation of motion

$$\mathrm{d}x_t = \mu_t^x x_t\,\mathrm{d}t + \sigma_t^x x_t\,\mathrm{d}z_t. \tag{10.5}$$

At this point, we do not say what x actually corresponds to. After formally writing down the agents' optimization problems and aggregating, we will see that the system can indeed be described by a single variable x – the experts' share of the total net worth.[27]

Assuming then, that the aggregate state is fully specified by x, it follows that its drift and diffusion coefficients are functions of x, $\mu_t^x = \mu^x(x_t)$, $\sigma_t^x = \sigma^x(x_t)$, and importantly, so is the the price process q:

$$q_t = q(x_t), \quad \mu_t^q = \mu^q(x_t), \quad \sigma_t^q = \sigma^q(x_t).$$

The Itô's Lemma allows us now to create a mapping from the aggregate state x to price q. Applying it to $q(x_t)$ yields

$$\mathrm{d}q_t = \left[\mu^x(x_t)x_t q'(x_t) + \frac{1}{2}\sigma^x(x_t)^2 x_t^2 q''(x_t)\right]\mathrm{d}t + \sigma^x(x_t)x_t q'(x_t)\,\mathrm{d}z_t. \tag{10.6}$$

Matching the drift and volatility terms in Eq. (10.6) with those from the original stochastic differential equation for the q process Eq. (10.3) yields the system of two equations:

$$\mu^q(x) = \mu^x(x)\frac{xq'(x)}{q(x)} + \frac{1}{2}\sigma^x(x)^2\frac{x^2q''(x)}{q(x)}, \tag{10.7a}$$

$$\sigma^q(x) = \sigma^x(x)\frac{xq'(x)}{q(x)}. \tag{10.7b}$$

Returning now to the agents' optimization problem, the controls consumption λ and asset allocation φ are to be chosen so as to maximize the objective function that now depends on the present agent net worth n ($\bar{n}$) and macro state x. In our example, we assume that agents have linear consumption preferences, so that an expert's value function is

$$V(n,x) = \max_{\varphi,\lambda} \quad \mathbb{E}\left[\int_0^\infty \mathrm{e}^{-\rho t}\lambda_t n_t\,\mathrm{d}t\right]. \tag{10.8}$$

The value function must satisfy the Hamilton-Jacobi-Bellman (HJB) equation [Technical Appendix A.2, Eq. (A.40)] which here reads

$$\begin{aligned}\rho V(n,x) = \max_{\lambda,\varphi}\Bigg\{&\lambda(n,x)n \\ &+\left[r + \left(\frac{a}{q(x)} + \mu^q(x) + \sigma\sigma^q(x) - r\right)\varphi(n,x) - \lambda(n,x)\right] n\frac{\partial V(n,x)}{\partial n} \\ &+\mu^x(x)x\frac{\partial V(n,x)}{\partial x} + \frac{1}{2}\sigma^x(x)^2x^2\frac{\partial^2 V(n,x)}{\partial x^2} \\ &+\sigma^x(x)x\left[\sigma + \sigma^q(x)\right]\varphi(n,x)n\frac{\partial^2 V(n,x)}{\partial n\partial x} \\ &+\frac{1}{2}\left[\sigma + \sigma^q(x)\right]^2\varphi(n,x)^2n^2\frac{\partial^2 V(n,x)}{\partial n^2}\Bigg\}.\end{aligned} \tag{10.9}$$

We cannot fix all boundary conditions for V at this stage, as we do not know what x is. Nonetheless, it is clear from Eq. (10.4) that if an agent has zero net worth, then n will always remain zero, as $\mathrm{d}n = 0$. Consumption will then also be zero, and so $V(0,x) = 0$ for all x. The objective function is linear in n, cf. Eq. (10.8), as are the n equations of motion, provided the controls are independent of n, and thus

$$V(n,x) = nW(x), \tag{10.10}$$

where $W(x)$ can be interpreted as the marginal value of net worth.

Substituting the factored V into Eq. (10.9), we reduce it to an ordinary differential equation that depends only on a single state variable x:

$$(\rho - r)W(x) = \max_{\lambda,\varphi}\Big\{\lambda(x)(1 - W(x))$$
$$+\varphi(x)\left[\frac{a}{q(x)} + \mu^q(x) + \sigma\sigma^q(x) - r + \sigma^x(x)(\sigma + \sigma^q(x))\frac{xW'(x)}{W(x)}\right]W(x)\Big\}$$
$$+\mu^x(x)xW'(x) + \frac{1}{2}[\sigma^x(x)x]^2W''(x). \tag{10.11}$$

It is easy to see that the right-hand side of (10.11) is linear in controls φ and λ. Thus, maximization in consumption λ implies

$$\lambda(x) = \begin{cases} 0, & \text{if } W(x) - 1 < 0, \\ \text{unbounded}, & \text{if otherwise.} \end{cases} \tag{10.12}$$

For households, the consumption $\bar{\lambda}$ choice is simpler: As they are not facing the non-negative consumption constraint, they choose their $\bar{\lambda}$ so that $\bar{W} = 1$.

If the coefficient of φ in Eq. (10.11) were positive, all experts would allocate an unbounded amount of their net worth to k (using infinite leverage to do so). As total k is constrained, the capital allocations must all be finite, which is consistent with the agents' optimization only if

$$\frac{a}{q(x)} + \mu^q(x) + \sigma\sigma^q(x) - r + \sigma^x(x)[\sigma + \sigma^q(x)]\frac{xW'(x)}{W(x)} = 0. \tag{10.13}$$

An equivalent formula holds for households and their capital allocation $\bar{\varphi}$, with the difference that they might prefer not to hold any capital at all:

$$\frac{\bar{a}}{q(x)} + \mu^q(x) + \sigma\sigma^q(x) - r \le 0, \quad \text{with equality if } \bar{\varphi} > 0. \tag{10.14}$$

Under (10.12) and (10.13), the expert HJB equation reduces to

$$(\rho - r)W(x) = \mu^x(x)xW'(x) + \frac{1}{2}[\sigma^x(x)x]^2W''(x), \tag{10.15}$$

for any value of φ and for all values of x such that $W(x) > 1$ holds. To fully close the model, one needs to pin down the equations of motion for the aggregate state – in other words, find and solve conditions determining diffusion coefficients $\mu^x(x)$ and $\sigma^x(x)$.

Noting that the drift and diffusion of expert (households) net worth is linear in n ($\bar{n}$), cf. Eq. (10.4), the total expert net worth, denoted N, follows

$$\mathrm{d}N_t = \left[r + \left(\frac{a}{q(x_t)} + \mu^q(x_t) + \sigma\sigma^q(x_t) - r\right)\varphi(x_t) + \lambda(x_t)\right]N_t\,\mathrm{d}t$$
$$+\left(\sigma + \sigma^q(x_t)\right)\varphi(x_t)N_t\,\mathrm{d}z_t. \tag{10.16}$$

Similar dynamics would emerge for total household net worth, $\bar{N}$. Now the aggregate state is determined by two state variables, N and $\bar{N}$ (x is of course still there, but here we are trying to identify what it should be). This reduces to one when one notes that debt and capital market clearing imply

$$N_t + \bar{N}_t = q_t K_t^{\text{tot}}, \tag{10.17}$$

where K_t^{tot} is the total capital in the economy. Aggregating the k equations motion the same way as was done above for n, we have that $\mathrm{d}K_t^{\text{tot}} = \sigma K_t^{\text{tot}} \mathrm{d}z_t$. We can now define the aggregate state variable to be the experts' share of the total net worth,

$$x_t \equiv \frac{N_t}{q_t K_t^{\text{tot}}}. \tag{10.18}$$

Itô differentiating the definition of x, we then have

$$\begin{aligned}\mathrm{d}x_t = \Big\{ \frac{a}{q(x_t)} \psi(x_t) + \Big[\mu^q(x_t) - \sigma^2 - \sigma\sigma^q(x_t) - \sigma^q(x_t)^2 - r\Big][\psi(x_t) - x_t] \\ - \lambda(x_t)x_t\Big\} \mathrm{d}t + [\sigma + \sigma^q(x_t)][\psi(x_t) - x_t]\,\mathrm{d}z_t, \quad (10.19)\end{aligned}$$

where ψ is the fraction of total capital held by the experts, $\psi(x) \equiv x\varphi(x)$. Equating the drift and diffusion terms of x as given by Eq. (10.19) with those coming from our earlier definition, Eq. (10.5), gives us what we will refer to as the *closure conditions*:

$$x\mu^x(x) = \frac{a}{q(x)} \psi(x) + \Big[\mu^q(x) - \sigma^2 - \sigma\sigma^q(x) - \sigma^q(x)^2 - r\Big][\psi(x) - x] \tag{10.20a}$$

$$x\sigma^x(x) = [\sigma + \sigma^q(x)][\psi(x) - x]. \tag{10.20b}$$

Finally, we can state the remaining boundary conditions for q and W. Experts will have unbounded consumption at the point where W reaches one, cf. Eq. (10.12). This introduces a reflecting upper boundary x^*, as whenever expert net worth share is over this point, they consume until x returns to the level x^*. By the properties of a reflecting boundary [Technical Appendix A.4.2], the derivatives at x^* must vanish, and we then have in total

$$q'(x^*) = 0, \qquad W'(x^*) = 0, \qquad W(x^*) = 1. \tag{10.21}$$

At the lower boundary, share of experts' net worth is stuck at zero, and so the price of capital there must be such that households are willing to hold it forever. As excess returns from holding capital for households are $\bar{a}/q(0) - r$ when price remains at $q(0)$, the least possible q must be $\bar{a}/r$. Finally, the marginal value of wealth $W(x)$ for experts must tend to infinity as $x \to 0$:[28] From Eq. (10.20b) we have that $\lim_{x\to 0} \varphi(x) = 1 + \sigma^x(0)/\sigma$. Assuming that experts are always leveraged,

$\varphi(x) > 1$, we must have that $\sigma^x(0) > 0$. Subtracting Eq. (10.14) from Eq. (10.13) we get that

$$\lim_{x\to 0} \frac{xW'(x)}{W(x)} = -\frac{a-\bar{a}}{q(0)\sigma^x(0)} < 0. \tag{10.22}$$

This implies that $W(x) \to \infty$ as $x \to 0$. Thus, at the lower boundary we have:

$$q(0) = \frac{\bar{a}}{r}, \qquad \lim_{x\to 0} W(x) = \infty. \tag{10.23}$$

In total, from Eqs. (10.21) and (10.23), we have five conditions, which is the correct number for two second-order ordinary differential equations, plus the yet unknown consumption boundary x^*.

We now have a sufficient number of equations to find the aggregate state drift and diffusion coefficients μ^x and σ^x, and the expert capital share ψ. In addition we should also state the differential equations determining W and q – we will obtain these when we solve our equations for W'' and q''. For these five unknowns, the five equations we need are the optimal capital allocation conditions, Eq. (10.13) and (10.14), the pair of closure conditions, Eqs. (10.20), and finally the expert HJB, Eq. (10.15). The solution is straight-forward, albeit the result is not particularly pretty:

$$\sigma^x(x) = \frac{a-\bar{a}}{q(x)}\left[-\frac{W(x)}{xW'(x)}\right]\left[\frac{\sigma}{2} + \sqrt{\left(\frac{\sigma}{2}\right)^2 - \frac{a-\bar{a}}{q(x)}\frac{xq'(x)}{q(x)}\frac{W(x)}{xW'(x)}}\right]^{-1}, \tag{10.24a}$$

$$\mu^x(x) = \frac{a}{q(x)} - \sigma^x(x)\left\{\sigma + \sigma^x(x)\left[\frac{xW'(x)}{W(x)} + \frac{xq'(x)}{q(x)}\right]\right\}, \tag{10.24b}$$

$$\psi(x) = x\left[1 + \frac{\sigma^x(x)}{\sigma + \sigma^x(x)\frac{xq'(x)}{q(x)}}\right], \tag{10.24c}$$

$$\begin{aligned} q''(x) = \frac{2q(x)}{x^2\sigma^x(x)^2}\Bigg\{ & r - \frac{a}{q(x)}\left(1 + \frac{xq'(x)}{q(x)}\right) \\ & + \sigma^x(x)\left[\sigma^x(x)\left(\frac{xq'(x)}{q(x)}\right)^2 - \sigma\frac{xW'(x)}{W(x)}\right]\Bigg\}, \end{aligned} \tag{10.24d}$$

$$W''(x) = \frac{2W(x)}{x^2\sigma^x(x)^2}\left[\rho - r - \mu^x(x)\frac{xW'(x)}{W(x)}\right]. \tag{10.24e}$$

The sign in front of the square root in Eq. (10.24a) is here chosen so that $\sigma^x(0) > 0$.

In the region where households do not hold capital, we have only four equations, as Eq. (10.14) used above becomes an inequality. On the other hand, we have only four unknowns as $\psi = 1$. Solving the remaining Eqs. (10.13, 10.20, 10.15), one now finds

$$\sigma^x(x) = \sigma \frac{1-x}{x-(1-x)\frac{xq'(x)}{q(x)}}, \tag{10.25a}$$

$$\mu^x(x) = \frac{a}{q(x)} - \frac{1-x}{x}\left[\sigma + \sigma^x(x)\frac{xq'(x)}{q(x)}\right]\left[\sigma + \sigma^x(x)\left(\frac{xq'(x)}{q(x)} + \frac{xW'(x)}{W(x)}\right)\right], \tag{10.25b}$$

$$\begin{aligned} q''(x) = \frac{2q(x)}{x^2\sigma^x(x)^2}\Bigg\{ & r - \frac{a}{q(x)} - (\sigma\sigma^x(x) + \mu^x(x))\frac{xq'(x)}{q(x)} \\ & - \left[\sigma + \sigma^x(x)\frac{xq'(x)}{q(x)}\right]\left[\sigma^x(x)\frac{xW'(x)}{W(x)}\right]\Bigg\}, \end{aligned} \tag{10.25c}$$

$$W''(x) = \frac{2W(x)}{x^2\sigma^x(x)^2}\left[\rho - r - \mu^x(x)\frac{xW'(x)}{W(x)}\right]. \tag{10.25d}$$

The model solution is complete once we numerically solve the q and W differential equations.

4.2 Numerical solution

4.2.1 Marginal value W and aggregate state x

The numerical solution of Eqs. (10.24) and (10.25) poses some challenges. Standard, local iterative ordinary differential equation solvers such as Runge-Kutta or predictor-corrector methods (see e.g., Hairer et al. (1993)) work by propagating the solution from a given point x forward or backward by evaluating the derivatives at and around x. In the case of Eqs. (10.24) and (10.25), such methods run into the problem of division by zero: At the left-hand side boundary, the derivatives of q and W evaluate to plus or minus infinity.

One solution to this problem is to ignore it: The solution of the equations can be attempted with the derivatives apparently evaluating to 1/0 or 0/0. Infinite initial values are replaced by very large, but finite numbers, and the derivatives are evaluated with the hope that numerical round-off error sends zeros to small but finite values, so that the undefined division by zero condition does not occur.

A more satisfactory approach is to remove the singularities altogether by some change of variables, or to use an approximate analytic solution near the critical point. In this example, we do the latter by constructing the asymptotic expansion of q and W near the boundary [Technical Appendix section A.5]. We begin by assuming a power law form for the solution near the lower boundary that is consistent with the boundary conditions of Eq. (10.23):

$$q(x) = \frac{\bar{a}}{r} + q_1 x^\alpha + o(x^\alpha), \qquad W(x) = W_1 x^{-\beta} + o(x^{-\beta}), \qquad \alpha, \beta > 0. \tag{10.26}$$

The exponents are determined by first substituting the trial functions into Eqs. (10.24d) and (10.24e), expanding the equations for small x, and then solving α and β so that the leading order term vanishes. Albeit the algebra is tedious, a solution eventually emerges:

$$\alpha = \frac{1}{2} - \beta\left(1 + \frac{a\bar{a}\beta\sigma^2}{(a-\bar{a})^2 r}\right) + \sqrt{\left[\frac{1}{2} - \beta\left(1 + \frac{a\bar{a}\beta\sigma^2}{(a-\bar{a})^2 r}\right)\right]^2 + \frac{2(\bar{a}\beta\sigma)^2}{(a-\bar{a})^2 r}}, \qquad (10.27a)$$

$$\beta = \frac{1}{2} - \frac{\bar{a}\rho}{2ar} - \frac{(a-\bar{a})^2 r}{4a\bar{a}\sigma^2} + \sqrt{\left[\frac{1}{2} - \frac{\bar{a}\rho}{2ar} - \frac{(a-\bar{a})^2 r}{4a\bar{a}\sigma^2}\right]^2 + \frac{(a-\bar{a})^2 r}{2a\bar{a}\sigma^2}}. \qquad (10.27b)$$

Four different combinations for the signs in front of the square roots are possible. Clearly, however, only the above choice yields solutions that are both positive. The coefficients q_1 and W_1 will be determined by the boundary conditions.

For small x, we can now use the trial solutions of Eq. (10.26), truncated to the displayed terms, with α and β from Eqs. (10.27). Then, for x greater than some small cross-over value ε, we use a standard iterative local ordinary differential equation solver, with initial conditions at ε coming from the asymptotic expansion. The value of ε should be large enough to ensure that evaluating the derivatives is not significantly affected by round-off error, but small enough so that the asymptotic expansion is accurate. As a first guess, the square root of the maximum relative error of the used floating point arithmetic, $\varepsilon \sim 10^{-8}$ for double precision, can be used.

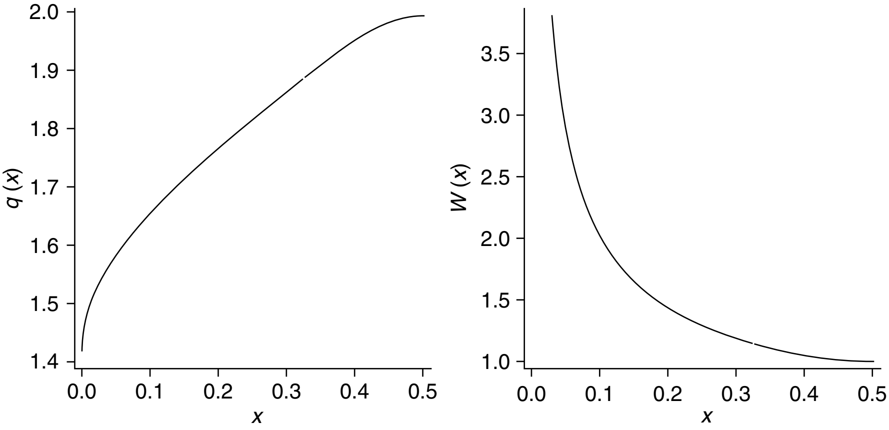
Using the above approach, we can now solve the q and W differential equations given q_1 and W_1, the coefficients of the leading non-constant terms in the asymptotic expansion of q and W (Eq. (10.26)). We can arbitrarily fix W_1 as the equations are invariant in linear scaling of W, and scale the solution ex-post in order to satisfy the condition $W(x^*) = 1$ at the consumption boundary. We still have q_1 and the position of the upper boundary x^* to be set so that the remaining boundary conditions $W'(x^*) = q'(x^*) = 0$ are satisfied.

A simple numerical scheme that finds q_1 and x^* can be set up as follows: Define $\Theta(q_1)$ as q' evaluated at first x such that $W'(x) = 0$ where q and W are solutions to the model equations for the given q_1. This point can be found by solving the differential equations forward from the initial point, until the $W'(x)$ boundary is crossed; the exact crossing point is then polished using standard root finding methods. The correct q_1 can then be determined by finding $\Theta(q_1) = 0$, where again, any standard root finding method can be employed.

In Figure 10.1 we have plotted the numerical solution following the method described above. The parameter values used are $a = 0.11$, $\bar{a} = 0.07$, $\sigma = 0.1$, $r = 0.05$, and $\rho = 0.06$.

4.2.2 Equilibrium probability distribution

The statistics of the possible realizations of the economy are given by the probability density function of x, $f(x)$ which itself is obtained from the Kolmogorov Forward equation [Technical Appendix section A.3, Eq. (A.45)]. In equilibrium,



(a) Capital price q as a function of the aggregate expert net worth share x.

(b) Marginal value of expert net worth as a function of the aggregate expert net worth share x.

Figure 10.1 Numerical solution of Eqs. (10.24) and (10.25)

for the process x of Eq. (10.19), this reads

$$0 = \mu^x(x)xf(x) - \frac{\partial}{\partial x}\left[\frac{1}{2}\sigma^x(x)^2x^2f(x)\right], \tag{10.28}$$

where the coefficients μ^x and σ^x come from Eqs. (10.24) or (10.25) depending whether households are holding capital or not, and where the q and W functions are presumed to have already been solved. Here, we have also used the fact that x^* is a reflecting boundary, and set the left-hand side of Eq. (10.28) to zero [Technical Appendix section A.4.2, Eq. (A.50)].

Rather than solving Eq. (10.28) directly, it is easier to define

$$f(x) = 2g(x)/(x\sigma^x(x))^2,$$

and solve for the function g instead. This change of variables avoids us having to differentiate σ^x, a straight-forward but laborious task. For g, the equation reads

$$0 = \frac{2\mu^x(x)}{x\sigma^x(x)^2}g(x) - g'(x). \tag{10.29}$$

The solution needs to be normalized to unit integral over the x range; to do this, we can solve the differential equation $F'(x) = f(x)$ in parallel to the one above. F will then be the cumulative probability distribution, if we further ask that $F(0) = 0$. Numerically, the equations can be solved with arbitrary initial conditions: If f and F are such un-normalized solutions, one simply replaces them according to $f(x) \mapsto f(x)/[F(x^*) - F(0)]$ and $F(x) \mapsto [F(x) - F(0)]/[F(x^*) - F(0)]$. This is valid since the f and F differential equations are invariant in scaling, the F equation also in the addition of constants.

As was the case with q and W equations, we have a singularity at $x = 0$, since $2\mu^x(x)/(x\sigma^x(x))$ tends to infinity at zero. An asymptotic expansion could be used here as well, and for completeness, we shall do it. But for the numerical solution, a simpler approach is possible: we can choose some interior point, and solve from there left towards the $x = 0$ boundary, and right up to $x = x^*$ edge. Due to the singularity, the left-hand side solution is likely to fail before reaching $x = 0$, but this is fine as long as we got near enough to 0, and the integral of density tends to a finite value (if not, in equilibrium all probability mass is at $x = 0$).

The $x \to 0$ asymptotic form of f can be found using the methods we have already used above. Alternatively, it would suffice to note that in the $x \to 0$ limits of the relative drift and diffusion of x are

$$\lim_{x\to 0} \sigma^x(x) = \frac{r}{\bar{a}\beta}\frac{a - \bar{a}}{\sigma}, \tag{10.30a}$$

$$\lim_{x\to 0} \mu^x(x) = \frac{r}{\bar{a}\beta}\left[\frac{r}{\bar{a}}\left(\frac{a - \bar{a}}{\sigma}\right)^2 + \bar{a} + (\beta - 1)a\right], \tag{10.30b}$$

and we can replace μ^x and σ^x by these limits in Eq. (10.28) and solve f analytically. The result is

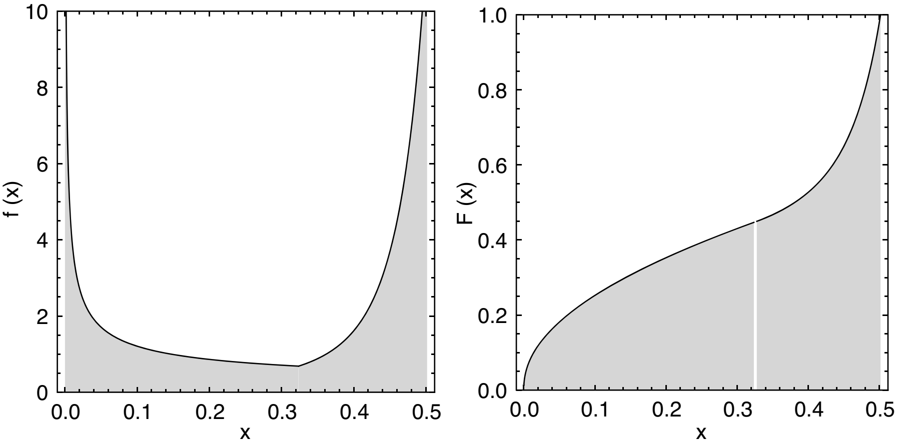
$$f(x) \propto x^\gamma + o(x^\gamma), \qquad \text{where } \gamma = 2\left[\beta - 1 + \beta\frac{\bar{a}(\bar{a} + (\beta - 1)a)\sigma^2}{(a - \bar{a})^2 r}\right]. \tag{10.31}$$

One immediately sees that if $\gamma < -1$, the integral of f is infinite over arbitrarily small interval $[0,\delta]$, $\delta > 0$. If this is the case, for those parameters, all probability condenses to $x = 0$.

The numerically solved probability distribution function f and the cumulative probability function F are plotted in Figure 10.2. Comparing to BS(2014-1), even in this simplified model, the two peaked structure of the density f is still visible. This is relatively unsurprising since the models differ mainly by the inclusion of investment dynamics. We refrain from analyzing the economic implications of the result, as the main goal was to present an easy example to understand derivation of the model and the solution methods.

5 Some paths for future research

We complete this chapter with a short discussion of the range of further issues that models of this kind might usefully address, emphasizing once again that this is still an immature literature, that many technical challenges remain and



(a) Probability density function f.

(b) Cumulative density function F.

Figure 10.2 Equilibrium probability distribution functions

therefore we can make no firm predictions about where contributions and breakthroughs will come.

There are many paths of future investigation that could be followed. A number of contributions to the new continuous-time macrofinancial literature assume an "AK" production function i.e., output is a linear function of the stock of capital. This assumption, which can be traced back to the contribution of Frankel (1962), was widely employed in the earlier contributions to the literature on endogenous growth (see Aghion and Howitt (2009) for review of this literature). Both Isohätälä et al. (2014) and Brunnermeier and Sannikov (2014a) can be interpreted as models of endogenous growth yielding the prediction that capital accumulation and hence the rate of economic growth, will fall when corporate net worth falls and that the economy may then remain for an extended period in a phase of low investment and low growth (the "net worth" trap). These models may thus already provide some insight into the major puzzle of the slowdown of aggregate productivity growth in many countries since the global financial crisis.

This is though a rather simplistic account of growth. The theoretical growth literature has moved on to focus on other mechanisms such as investment in product variety (as in Romer (1987, 1990)) and in the discovery of new more efficient methods of production (innovation) that replace older inferior methods (e.g., the model of Schumpterian "creative destruction" of Aghion and Howitt (1992)). A natural further development will therefore be to employ similar continuous time models to examine the impact of balance sheet constraints and financial distress on investment in new products and processes and hence on productivity growth. Doing this though may be difficult within the current assumption of only a single state variable.

A related issue is that of structural adjustment following financial crises. Many countries need to adjust the structure of their economies, for example switching labor and capital resources from non-traded to traded output. Rising risk-premia in periods of financial distress can act as a barrier to such investments, providing another form of trap in a low-output low-income state. Other similar issues arise in understanding the low elasticities of traded sector output to changes in exchange rates following financial crises or in the response of small open economies to "sudden stops" of capital flows (on this issue Brunnermeier and Sannikov (2015) have made a promising start using continuous time methods).

Further understanding will surely also need to take account of the interaction of balance sheet constraints with household, corporate and bank expectations about future productivity and incomes. To date the continuous-time macrofinancial literature has imposed the conventional but rather strong assumption of model-consistent expectations. Every agent is assumed to know both the current state of the economy and stochastic processes that drives both the state of the economy and market prices. Different, possibly even more extended, dynamics can be expected when agents update their expectations about unobserved states and processes in response to their current observations. Thus, for example, a fully adequate model of endogenously created bank inside-money would seem to need to take account of the possibility that optimistic expectations about future income result in a period of rapid and self-reinforcing expansion of both bank credit and bank money. Introducing learning of this kind into these models will be a further technical challenge.

A further potential line of inquiry is to make the treatment of financial markets more realistic, for example by allowing for alternative (financial markets) source of financing for the productive sector, on top of bank loans, since the empirical evidence suggests that firms tend to partly substitute bank financing by market financing when credit conditions tighten (see e.g., Becker and Ivashina (2014)). This would help to get a better understanding of how the substitutability of funding sources can add to/mitigate the propagation of fundamental shocks, with implications for growth and financial stability. Another similar departure would be to allow for the internal source of financing for the productive sector (i.e., capital accumulation within firms). Introducing this feature, most likely, would require introducing an additional state variable in play – the net worth of the productive sector, which might be technically challenging.[29]

Similarly there is need for better understanding of the role of both commercial and central bank balance sheets in macroeconomic transmission and the supply of credit. The DSGE assumption that all that matters in monetary policy is interest rates is now accepted as an oversimplification, but we are not yet in the position of having tractable incomplete market models, in either continuous or discrete time, in which the role of commercial and central bank balance sheets is clearly articulated. While the work of Brunnermeier and Sannikov (2014d) that we have reviewed offers a particularly promising start, it is still not yet possible to say that the present "state of the art" is sufficiently developed to provide a full understanding of the impact of unorthodox monetary policy (central bank balance sheet expansion) or macroprudential tools (such as cyclically varying capital requirements or limits on loan-to value ratios).

These paths for future research are far from exhausting the list of possible applications of new continuous-time macrofinancial modeling. There are also opportunities to apply this framework in a number of other settings more routinely explored using standard linearized DSGE models. Examples include modeling the labor market, real wages and employment and product markets and price setting. If balance sheet constraints affect investment and asset markets then they should also affect labor and goods markets. Writing down such models in continuous time with balance sheet and net worth constraints does not seem so difficult. Solution though could be challenging because of the need to include additional state variables.

The challenges of numerical solutions should give pause for thinking carefully about the choice of modeling strategy. There are well developed tools for the numerical solution of macroeconomic models with several state variables in discrete time. Replicating all this technical work on solutions with many state variables in continuous time may not be an efficient way to proceed. It may instead be more useful to find ways to incorporate the insights of continuous-time macrofinancial modeling into more widely known and understood discrete time settings. This is one reason why, as we have already suggested, we believe that over time the "gap" between continuous time and discrete time specifications can and should be closed. There is no reason why the impact of balance sheet constraints and net worth cannot be incorporated into otherwise standard discrete-time specifications (although this may come at some cost, for example the need to introduce more explicit modeling of what happens when disturbances result in constraints binding, something that can often be conveniently put to one side when uncertainty is modeled as a continuous time diffusion).

Finally, of course, it will be essential to take these models closer to data. Some initial steps in this direction are already made by He and Krishnamurthy (2013, 2014) who seek to replicate the asset market's behavior during the 2007–2009 financial crisis. Improving predictive and simulation properties of these models, however, carries a risk of losing tractability and transparency. An illustration is the challenge of explaining the counterfactual prediction of He and Krishnamurthy (2012) and He and Krishnamurthy (2013), that the leverage of financial intermediaries increases substantially during financial crises (the underlying mechanism is that the emergence of crisis results from low specialist equity requiring them to leverage in order to maintain "skin in the game").

As an example of the challenges of bringing these models to the data, we can briefly describe how the model by Adrian and Boyarchenko (2013) employs continuous-time macrofinancial modeling tools to develop an explanation of observed pro-cyclical intermediary leverage. There are several distinctive features of this work. Three of these seem to be particularly important. First that financial intermediaries do not maximize an objective function (such as present discounted future dividends); instead, they behave mechanically, first using earnings to pay floating rate coupons on long term bonds issued to households (the coupon rate is determined by equilibrium of the supply and demand for these bonds) and retaining all remaining earnings to build up equity and invest in productive capital. Second that if intermediary equity falls below a lower boundary, then financial intermediaries are restructured, with debt-holders wiped out and re-established under new equity holders. Since all intermediaries are identical and hit by the same shocks this is a systemic crisis. Third that financial intermediary leverage, and hence their investment in productive capital, is continuously determined by a regulatory driven, value-at-risk-type constraint that depends on the short term volatility of the price of productive capital.

Similar to the other macrofinancial models that we have described, all variables of interest (intermediary equity, the price of productive capital, the expected excess returns on holding intermediaries' debt and productive capital) can be described as the functions of a single state variable. Just as in He and Krishnamurthy (2012) and Brunnermeier and Sannikov (2014a) this state variable is the share of financial intermediaries' net worth in total wealth. The novel contribution is two closely related empirical predictions not captured by other models. First, as a direct consequence of the assumed leverage constraint, this model it generates the empirically-observed pro-cyclical pattern of intermediary leverage. Second it also explains what Adrian and Boyarchenko (2013) describe as the volatility paradox, i.e., the well known observation that systemic risks tend to increase during periods of perceived low volatility: e.g., for example, during the "great moderation" that preceded the global financial crisis. In their model this appears as a negative relationship between the instantaneous endogenous volatility of the returns to holding capital with the probability of a systemic default. As endogenous volatility declines, leverage rises and thus also does the risk of a systemic default on a six-months-ahead horizon (see their figure 5.)

While Adrian and Boyarchenko (2013) make a valuable further contribution, their work can also be read as an illustration of the very substantial challenges of bringing models of this kind to the data. Departures from forward looking behavior or intertemporal optimization are not necessarily wrong, but these departures open up such a large menu of possible modeling choices that is difficult to know what is the best way forward. There is an almost unlimited range of possible underlying assumptions of this kind that can generate macro-financial interactions.

In order to impose the required intellectual discipline, it may prove necessary to focus on developing models of incomplete markets that are consistent, not just with observed aggregate outcomes such as asset prices or national accounting measures of output and investment (there are simply too many potential modeling choices for doing this), but also with micro level data at the level of individual firms, households and financial institutions. In this context it will be difficult to ignore the simplifying aggregation assumptions employed in the continuous-time macrofinancial literature. Instead it may eventually prove necessary to work with large scale agent-based models, in which distribution of net worth and leverage within sectors is tracked as well as the aggregate net worth (see Haldane (2015) for further discussion of why macrofinancial modeling should use agent-based approaches). Work of this kind will however need a different approach to research than has been conventionally used in macroeconomics, requiring relatively large teams of researchers in order to collect and match the underlying microlevel data.

6 Conclusion

This chapter has reviewed several contributions to a new and promising current literature, employing continuous time models to capture some of the macrofinancial interactions that have been highlighted by the global financial crisis. Though using highly stylized specifications, these models demonstrate how the interaction of market incompleteness with the balance sheet constraints of economic agents can generate dynamics of macroeconomic variables much more consistent with the empirically observed patterns at times of crisis than those generated by conventional DSGE models. These dynamics include substantial variations in risk-premia and asset prices, as well as subsequent substantial and highly persistent declines of macroeconomic aggregates such as output and investment. In the models reviewed in this chapter these dynamics are largely driven by what are in effect changes in attitudes to risk, and by externalities stemming from the fact that individual agents do not internalize the impact of their individual risk taking or other decisions on the wider economy.

As we highlight in Section 5 of this chapter, the modeling approach introduced by this new literature has the potential to address a large spectrum of macroeconomic problems. The pursuit of these avenues of future research is an exciting challenges, from both technical and economic perspectives. Where then will future research ultimately take us? We have no crystal ball but our judgment is that the eventual destination of this literature will be a relatively small number of comparatively simple but influential continuous-time models of the kind we have reviewed here, providing widely accepted economic intuition and policy insight into macrofinancial interactions rather than predicting accurately macroeconomic and financial market developments.

An important impact of these modeling efforts may be persuading researchers working in more conventional discrete-time frameworks of the necessity of taking incomplete markets and balance sheet and net worth constraints seriously. Matching with data is then in turn likely to require more “agent-based” approaches in order to meet the fundamental challenge of aggregation.

Such research will need a very careful process of matching against both microlevel and aggregate data in order to develop useful models. The final outcome could be a shift in modeling paradigms, from the typical small-team work found in much current macro economics to more resource intensive investigations involving many investigators and massive efforts at data-collection and calibration, with the overall direction of research guided to an important degree by the insights of the new continuous time approach to macrofinancial modeling.

This does not avoid the need for considerable efforts in order to find a reasonable balance between, on the one hand transparency and clear economic intuition, and on the other the realism and accuracy of underlying economic assumptions. We believe that the new macrofinancal models, despite the use of techniques of continuous-time stochastic modeling with which most economists are not very familiar, are especially useful because of the relatively clear and simple intuitions they provide about the macroeconomic consequences of incomplete markets, and hence the resulting impact of balance sheet and net worth on macroeconomic outcomes. We therefore hope this chapter can be helpful in acquainting our readers with the conceptual and technical features of this new generation of models and stimulating interest in this field of research.

A Appendix: Basics of continuous time models

In this appendix, we review some technical issues that arise when trying to understand, construct, or solve models that incorporate binding financing constraints in continuous time models. We begin with a brief summary of key concepts, eschewing formal proofs and favoring heuristic derivations that nonetheless can in principle be used as a basis for a more rigorous approach.

A.1 Stochastic differential equations

The primary modeling tool for describing the state of an agent is the *stochastic differential equation* (SDE), a generalization of the ordinary differential equation that incorporates some form of external stochastic forcing. Gardiner (2009) gives an excellent and approachable review of many of the topics covered here and subsequent sections; other solid reference texts on stochastic differential equations include Øksendal (2003) and Feller (1971).

Let x_t stand for an agent's state, e.g., wealth or net worth, at time t, with the initial, $t = 0$ state x_0 given. We say x is a *diffusion process* and formally write its equation of motion as the SDE

$$\mathrm{d}x_t = \mu(x_t)\mathrm{d}t + \sigma(x_t)\mathrm{d}z_t, \tag{A.32}$$

where z_t is a *Wiener process*, a continuous stochastic process having independent and normally distributed increments, $z_{t+h} - z_t = \sqrt{h}\eta$, where η is a unit normal distributed random variable. The differentials, $\mathrm{d}x_t, \mathrm{d}z_t, \ldots$, can be viewed as the zero time step limits of corresponding finite differences, so that Eq. (A.32) becomes the limit of the discrete time model

$$x_{(k+1)T} = x_{kT} + T\mu(x_{kT}) + \sqrt{T}\sigma(x_{kT})\eta_{kT}, \tag{A.33}$$

where T, $T \to 0$, is the length of the period and η_{kT}, $k = 0, 1, \ldots$, are independent unit normal distributed random variables.[30] More rigorously, the stochastic differential equation, Eq. (A.32) should be understood as a short-hand way of writing the *stochastic integral*

$$x_t = x_0 + \int_0^t \mu(x_s)\mathrm{d}s + \int_0^t \sigma(x_s)\mathrm{d}z_s, \tag{A.34}$$

where the integral against the Wiener process, or any diffusion process in general, is defined analogously to the Riemann-Stieltjes integral. *Stochastic calculus* is the theory of stochastic differential equations; for a more focused reference on the topic, we refer the reader to e.g., Klebaner (2005).

An important analytical tool is *Itô's Lemma* which allows one to differentiate functions of diffusion processes such as x as given by Eq. (A.32). If f is any twice differentiable function defined in the domain of x, then f evaluated at x_t, denoted $f_t = f(x_t)$, is also a diffusion process with the increments

$$\mathrm{d}f_t = \left[\frac{\partial f(x_t)}{\partial t} + \mu(x_t)\frac{\partial f(x_t)}{\partial x} + \frac{1}{2}\sigma(x_t)^2\frac{\partial^2 f(x_t)}{\partial x^2}\right]\mathrm{d}t + \frac{\partial f(x_t)}{\partial x}\sigma(x_t)\mathrm{d}z_t. \tag{A.35}$$

Although we focus on single state variable problems, equivalent equations for multivariate case are still useful. For instance, in the construction of our example model, we needed multivariable formulas as we reduced an initially two variable problem to a single variable. Suppose X_t is an N-dimensional diffusion process taking values on some subset of $\mathbb{R}^N$, $X_t = (x_t^1, x_t^2, \ldots, x_t^N)^\mathsf{T}$ (here $^\mathsf{T}$ stands for matrix transposition). We write the X equation of motion as

$$\mathrm{d}X_t = \mu(X_t)\mathrm{d}t + \sigma(X_t)\mathrm{d}Z_t, \tag{A.36}$$

where now Z is a vector of Wiener processes, $Z_t = (z_t^1, \dots, z_t^K)^\mathsf{T}$, K is the number of independent shock sources, $\mu(X_t) = (\mu^1(X_t), \dots, \mu^N(X_t))^\mathsf{T}$ is the drift vector and $\sigma(X_t) = [\sigma^{ij}(X_t)]_{ij}$ is the $N \times K$ covariance matrix. The elements of Z_t can be without loss of generality taken to be independent: any and all instantaneous correlations between shocks are encoded in the matrix σ.

The multivariate version of Itô's Lemma, Eq. (A.35), for a function $f = f(X)$, $f : \mathbb{R}^N \mapsto \mathbb{R}$, reads

$$\mathrm{d}f_t = \left\{ \mu(X_t)^\mathsf{T} \nabla_X f(X_t) + \frac{1}{2} \operatorname{tr} \left[\sigma(X_t)^\mathsf{T} \mathsf{H}_X f(X_t) \sigma(X_t) \right] \right\} \mathrm{d}t + \nabla_X f(X_t)^\mathsf{T} \sigma(X_t) \, \mathrm{d}Z_t, \quad \text{(A.37)}$$

where tr stands for matrix trace, and $\nabla_X f$ and $\mathsf{H}_X f$ give the gradient vector and the $N \times N$ Hessian matrix of the function f, $[\nabla_X f(X)]_i = \partial f(X)/\partial x^i$, $[\mathsf{H}_X f(X)]_{ij} = \partial^2 f(X)/\partial x^i \partial x^j$.

These equations assume that the SDEs do not depend explicitly on time t; that is, they are time-homogeneous. An easy way of extending all of these definitions to account for explicit time dependence is to consider t additional state variable in X, with the trivial SDE $\mathrm{d}t = \mathrm{d}t$.

A.2 Hamilton-Jacobi-Bellman equation

The standard tool in stochastic dynamical programing is the *Hamilton-Jacobi-Bellman equation* (Fleming and Soner, 2006). Suppose that equations of motion, Eqs. (A.32) and (A.36) in the multivariate case depend on some controls y, which we now wish to choose so that our discounted future utility is maximal. For an infinite time-horizon problem, with standard exponential discounting and time-preference rate ρ, the objective function to maximize is

$$\Omega(x_0; \{y_t\}_{t=0}^\infty) = \mathbb{E} \int_0^\infty \mathrm{e}^{-\rho t} u(y_t) \, \mathrm{d}t, \quad \text{(A.38)}$$

where u is the utility function. Let V then be the maximal Ω, and henceforth assume that y refers to the maximizer:

$$V(x) = \max_{\{y_t\}_{t=0}^\infty} \Omega(x_0; \{y_t\}_{t=0}^\infty). \quad \text{(A.39)}$$

Assuming that V is twice differentiable, it can be found as a solution to the Hamilton-Jacobi-Bellman (HJB) equation:

$$\rho V(x) = \max_y \left\{ u(y) + \mu(x,y) V'(x) + \frac{1}{2} \sigma(x,y)^2 V''(x) \right\}. \quad \text{(A.40)}$$

An easy heuristic derivation goes as follows: (*i*) Start with the definition of V and divide the integral into two parts: from 0 to some small h, and from h to infinity; (*ii*) Use Bellman's principle on the second integral to make it V a time

h later, appropriately discounted, $e^{-\rho h}V(x_h)$; (*iii*) Use Itô's Lemma to approximate $V(x_h)$, Taylor expand in h, and then let $h \to 0$. This basic derivation of the Hamilton-Jacobi-Bellman equation requires that V be twice differentiable. It is well known that solutions do not always have this property. The theory of viscosity solutions addresses this problem, however, this topic is beyond the scope of this introduction (See e.g., Fleming and Soner (2006), or Crandall et al. (1992) for a rigorous but self-contained guide to the subject).

Generalizing to the multivariate case, the Hamilton-Jacobi-Bellman equation for the process of Eq. (A.36) with controls y reads:

$$\rho V(X) = \max_{y}\left\{u(y) + \mu(X)^{\mathsf{T}}\nabla_X V(X) + \frac{1}{2}\operatorname{tr}\left[\sigma(X)^{\mathsf{T}}\mathsf{H}_X V(X)\sigma(X)\right]\right\}. \tag{A.41}$$

A.3 Equilibrium characterization

In the conventional view, persistent e.g., in much of DSGE modeling today, what constitutes an equilibrium is a point in state space, plus random fluctuations induced by shocks. Such equilibria are a feature of, say, models linearized around a deterministic steady state, and which treat shocks as relatively small perturbations. When the modeling paradigm allows for large deviations, as is the case in the models we are highlighting here, this point-plus-perturbations picture of the equilibrium breaks down. Shocks, possibly amplified by feedback effects, can now drive the system far from what would have traditionally been seen as a relatively tranquil equilibrium point. Rather then, the equilibrium is characterized by a *probability distribution over the whole state space*.

In continuous time, the probability distribution of a diffusion process x is given by the *Kolmogorov forward equation*, also known as the *Fokker-Planck equation* (See e.g., Gardiner (2009); Risken (1996) is solely dedicated this equation): If $f(t,x|x_0)$ is the probability density function of process x following Eq. (A.32) with initial data x_0, then f satisfies the partial differential equation (omitting the explicit conditioning on the initial x):

$$\frac{\partial f}{\partial t}(t,x) = -\frac{\partial}{\partial x}\left\{\mu(x)f(t,x) - \frac{\partial}{\partial x}\left[\frac{1}{2}\sigma(x)^2 f(t,x)\right]\right\}. \tag{A.42}$$

Probability densities are integrated to get actual probabilities: Given a large number of independent realizations of x, the probability of finding x in the interval $[x_0, x_1]$ at time t is

$$\mathbb{P}_t(x_t \in [x, x+\Delta x]) = \int_{x_0}^{x_1} f(t,x')\,\mathrm{d}x'. \tag{A.43}$$

The "large number of independent realizations" can be understood either as many simultaneously running independent processes (and so with independent shocks), or as a large number of samples of a single process, taken over an infinitely long time period. In the former view, $f(x)$ represents the cross-sectional density of the state variables following the same dynamic stochastic equations of motion. Which view is correct depends of course on what one aims to model.

For additional intuition, the forward equation can be written in the form of a continuity equation relating the temporal change of f to spatial variation of a probability flux:

$$\frac{\partial f}{\partial t}(t,x) = -\frac{\partial j}{\partial x}(t,x), \tag{A.44a}$$

$$j(t,x) = \mu(x)f(t,x) - \frac{\partial}{\partial x}\left[\frac{1}{2}\sigma(x)^2 f(t,x)\right], \tag{A.44b}$$

where $j(t,x)$ is the probability current; that is, the rate of flow of probability through the point x to the positive x direction. In equilibrium $\partial f(t,x)/\partial t = 0$, and therefore the Fokker-Planck equation reduces to a first order ordinary differential equation

$$\mu(x)f(t,x) - \frac{\partial}{\partial x}\left[\frac{1}{2}\sigma(x)^2 f(t,x)\right] = j_0, \tag{A.45}$$

where j_0 is a constant to be determined by the boundary conditions.

The multivariate Fokker-Planck equation corresponding to the process X, again taking values on $\mathbb{R}^N$ and following the SDE (A.36), is in turn

$$\frac{\partial f}{\partial t}(t,X) = -\nabla_X \cdot J(t,X) \tag{A.46a}$$

$$J(t,X) = \mu(X)f(t,X) - \nabla_X \cdot \left[\frac{1}{2}\sigma(x)^{\mathsf{T}}\sigma(x)f(t,X)\right], \tag{A.46b}$$

where $J = (j^1,\ldots,j^N)^{\mathsf{T}}$ is now an N-dimensional probability current, and $\nabla_X\cdot$ is the divergence operator, $\nabla_X \cdot J(X) = \sum_{i=1}^N \partial j^i(X)/\partial x^i$, $[\nabla_X \cdot \sigma(x)^{\mathsf{T}}\sigma(X)]_i = \sum_{j=1}^N \partial[\sigma(x)^{\mathsf{T}}\sigma(x)]_{ij}/\partial x^j$.

A.4 Boundary conditions

Solutions to the HJB and the Fokker-Planck equations, Eqs. (A.40) and (A.44), or Eqs. (A.41) and (A.46) in the multivariate case, are not uniquely fixed until appropriate boundary conditions are given. Nonlinearity, capital constraints, and the need for proper treatment boundary conditions go hand in hand: If one is to construct a model that can account for large fluctuations, one must account for the possibility of a state variable hitting a hard bound, e.g., a capital or leverage constraint. Here, we consider the two most common boundary conditions: the absorbing and the (instantaneously) reflecting boundary.

A.4.1 Absorbing boundary

An absorbing boundary, say placed at position $x^\dagger$, is such that upon reaching it, the process is stopped and removed from the distribution. A stopped process (an agent who goes out of business, or a firm that has been liquidated) can no longer generate utility, and so it is natural to require that at an absorbing boundary the value is zero,

$$V(x^{\dagger}) = 0, \tag{A.47}$$

whenever of course the utility function is non-negative for all controls. For the probability density, an absorbing boundary at $x^{\dagger}$ means that

$$f(x^{\dagger}) = 0, \tag{A.48}$$

which has the natural interpretation of asking that point $x^{\dagger}$ is always completely free of the process x.

Multivariate generalizations are obvious: The absorbing boundary is not a point anymore, but some surface in the embedding space, and the same zero value or zero density requirement holds.

A.4.2 Reflecting boundary

Although the word “reflection” invokes a picture of a very certain type of motion, such as the elastic bouncing of a ball off of a rigid wall, or specular reflection of a beam of light, a reflecting boundary is here understood somewhat more generally. We will say that a boundary is reflecting whenever it conserves probabilities in the sense that it does not leak probability in or out, or allow the process to accumulate or to stop there for a finite time period.

In general, a reflecting boundary is set up by some forcing term that is strong enough to overcome the drift and diffusion terms in Eq. (A.32), preventing the process from ever crossing the boundary. Such forcing can be due to e.g., a singular control term (a control that has unbounded magnitude and which optimally is always either fully on or off). A rigorous mathematical treatment of SDEs with reflection does not use infinitely strong drift terms,[31] but as a model to guide intuition, the idea that the boundary is enforced by infinitely strong and infinitely short kicks is reasonable enough.

For functions of process x, a reflecting boundary implies a Neumann condition: it imposes a specific value on the derivative of the function. Say there is a reflecting boundary at x^*. For the value function, we then must have that

$$V'(x^*) = 0. \tag{A.49}$$

This can be justified as follows: imagine the x range inside the boundary reflected into the range outside the boundary. One can now view a process hitting the boundary as instead passing into the “mirror” space. In order for V to be smooth across the boundary, demanded by the smoothness of optimal V, then the derivative V' is zero. Similarly, for any continuously differentiable function of the process, the derivative should vanish at x^*.

For the density f, a reflecting boundary naturally corresponds to a point where the probability current j, Eq. (A.44), vanishes

$$j(x^*) = \mu(x)f(x) - \frac{\partial}{\partial x}\frac{1}{2}\sigma(x)^2 f(x)\bigg|_{x=x^*} = j_0 = 0. \tag{A.50}$$

Note that in the one-dimensional case, in steady state, this condition fixes the probability flow to zero over the whole of the x range. If there is also a reachable absorbing state, the probability density must be over the x range, for then both the value and derivative of f vanish at the same time. The word reachable is key: An absorbing boundary may be such that it cannot be arrived at in finite time. In this case, both an absorbing and reflecting boundary can co-exist, with the probability density not collapsing to zero.

Extensions to N-dimensional processes X are somewhat more complicated than for the absorbing boundary. Suppose that the reflecting boundary is a surface in $\mathbb{R}^N$, x^* is a point on that surface, and that $\Gamma(x^*)$ is the direction of the boundary forcing term (assumed never perpendicular to the normal of the boundary). Then the boundary condition for V reads

$$\Gamma(x^*) \cdot \nabla_X V(x^*) = 0, \tag{A.51}$$

that is, the Γ-directed derivative of V is zero when on a reflecting boundary.

For the Fokker-Planck equation, the N-variable extension Eq. (A.50) is

$$\nu(x^*) \cdot J(x^*) = 0, \tag{A.52}$$

where $\nu(x^*)$ is the inwards unit normal vector of the boundary surface at x^*, and J is the probability current as given by Eq. (A.46b). The natural interpretation is that the probability flow perpendicular to the surface is zero (no outflow of probability, or accumulation on the surface).

A.5 Asymptotic analysis

A useful tool in studying the behavior of continuous time models is the asymptotic expansion. These are simply approximate analytical solutions of the model equations that are valid only near the boundaries. Their utility lies in the fact that they can yield analytic insight into the qualitative and quantitative behavior of the model near the boundaries, which in turn can aid the model analysis or help with the numerical solution of the equations.

The exact way of constructing the expansion varies from problem to problem, but the general idea is to use the smallness of the distance to the boundary, or the greatness of the variable if very far from it, as a simplifying assumption. A fairly generally applicable recipe goes as follows:

1. Guess the limiting form of the solution, oftentimes a power law of the independent variable.

2. Substitute this trial function into the equation to be solved.
3. Expand the equation to leading order by neglecting terms that are guaranteed to be smaller than other terms in the equation.
4. Choose the parameters of the trial function so that a solution matching the boundary constraints are satisfied.

Notes

1. An example of such a DSGE extension is Meh and Moran (2010) who generalize the financial accelerator to include a bank-moral hazard based on Holmstrom and Tirole (1997).
2. Our paper complements Brunnermeier and Sannikov (2016) who provide a detailed discussion of the solution methods employed in these continuous-time models of this kind.
3. Generally, these are partial differential equations, but when the model in question has just a single state variable, as is the case in the models we review here, the equations become ordinary differential equations.
4. See Guvenen (2011) for a detailed review of this literature.
5. The standard results are those of Gorman (1959), who considers restrictions on utility under which consumption of goods can be expressed as a linear function of wealth allowing the choices of a large number of households to be restated as that of a representative consumer; and of Rubinstein (1974) and Constantinides (1982) who examine aggregation in the context of portfolio allocation-consumption decisions. Constantinides (1982) shows that under relatively weak conditions with complete financial markets the decisions of individual consumers can be replaced by that of a composite representative agent. See Guvenen (2011) for more discussion.
6. Similarly strong representative agent assumptions are also imposed in earlier literature on the macroeconomics of financial frictions, including in the influential work of Kiyotaki and Moore (1997) and Bernanke et al. (1999).
7. The Krusell-Smith algorithm for obtaining model consistent capital dynamics is based on updating a linear rule for the period by period investment in the stock of capital through a regression on the simulated model output from the previous iteration. Iteration continues until the investment rule is model consistent and the accuracy of the numerical solution is judged by the fit of the regression.
8. For further discussion see Den Haan (2010).
9. Another way of thinking about these challenges of numerical convergence is that an algorithm of this kind in effect substitutes moments of the distribution of networth, both across individual agents and across time, for the full distribution. If insufficient moments are included then the algorithm may yield a poor approximation to the correct solution.
10. In Section 4 of this chapter we discuss the technicalities of solution of a simple illustrative example of continuous time macrofinancial modeling, hoping in this way to make this literature accessible to readers who are much more familiar with discrete time modeling. We also recommend as good practice further steps to help readers become acquainted with these methods. One helpful presentational device, used for example by Klimenko et al. (2015), is to first state a model in discrete time with time steps of length Δt and then derive the limit as $\Delta t \to 0$. Another helpful step is to develop standalone numerical solvers which allow readers to use "sliders" to vary parameters and observe the consequent changes in solutions. The

website www.leveragecycles.lboro.ac.uk contains examples of such standalone solution software for two of the papers reviewed here, Isohätälä et al. (2014) and Isohätälä et al..

11. This, however, does not apply to *jump*-diffusion processes.
12. See Krueger and Kubler (2008) for a short overview, including discussion of the challenge of computing solution in a small number of state variables when it is no longer possible to obtain solution using contraction mapping theorems (theorems closely related to the aggregation results of Constantinides (1982) and the implied correspondence between market equilibrium and an equivalent central planning problem). The algorithm of Krusell and Smith (1998) is the most widely cited example of such methods applied in the context of incomplete markets. Ljungqvist and Sargent (2000) chapter 17 offer a number of other examples of solutions for incomplete market economies and Feng et al. (2014) and Guerrieri and Iacoviello (2015) for two recent proposed methods for recursive numerical solution of incomplete market models in discrete time. Tractable solutions of these models are described as "Markovian" because the stochastic dynamics can be expressed in terms of the equations of motion of a limited number of state variables.
13. By contrast, the models by Brunnermeier and Sannikov (2014b) and Brunnermeier and Sannikov (2014d) that we review below enable the asset to be traded among two classes of agents, which allows capturing the impact of "fire sales" on asset prices and track their feedback into the dynamics of agents' wealth.
14. A further assumption, introduced in order to avoid the challenges of solving for punishment and reward strategies as a dynamic game, is that the contract between households and specialists lasts only from t to $t + \Delta t$ after which the relationship between household and specialist is broken and each household is paired with a new specialist. This means that the equity constraint emerges as the solution to a static bargaining problem.
15. This property emerges essentially due to the absence of leverage in the unconstrained region. In the models we review next, the endogenous volatility is affected by the changes in leverage/feedbacks from asset prices and does not remain constant even when the capital/leverage constraints are far from binding.
16. Note that households are no longer infinitely lived. Instead, HK(2013) consider the continuous time limit of an "overlapping generations" setting in which households, born and then die almost instantaneously. Specifically, households are born at t with a labor income proportional to the dividend on risky assets, and allocated in proportion $\lambda : 1 - \lambda$ to one of two classes of "risk-averse households" whose wealth must all be held in the form of loans to specialists, and "risk-tolerant" households who are free to choose the proportion of their wealth invested in risky assets, managed by specialists, and in loans to specialists. Households consume at t in order to maximize a utility function log linear in current consumption and an end-period bequest at $t + \Delta t$ randomly allocated across the next generation (labor income is of infinitesimal size relative to inherited wealth and utility is logarithmic, implying that household consumption is a fixed proportion of their inherited wealth, the random allocation avoids the necessity of tracking the distributional impact of the allocation to risk-averse and risk-tolerant classes).
17. The title of their paper "A macroeconomic model with a financial sector" needs some explanation. Their productive experts who engage in investment and production could be real economy firms but on this interpretation their model does not have a financial sector at all; the title reflects their assumption that the assets held by these firms can be freely bought and sold between experts and households suggesting that they actually have in mind a very similar setting to that of HK(2012) and HK(2013) and that their experts are financial intermediaries who manage tradeable assets (see

Brunnermeier and Sannikov (2014b), the online appendix to Brunnermeier and Sannikov (2014a), where an equivalent version of their model distinguishing financial intermediaries and productive firms is discussed).

18. In BS(2014-1) setting experts do not need to maintain any liquid reserves, as arises in structural corporate finance models in which there are costs of adjusting liabilities (see e.g., Bolton et al. (2011)).
19. As shown in our illustration in Section 4, this feature is not crucial. Aside from the investment impact, the principal model results hold when this channel is switched off.
20. By contrast, if experts could costlessly issue new equity, there would be no capital traded and all capital would instead be held by experts. The price of capital then would be constant and would reflect the expected discounted value of the perpetual output stream under the more productive technology.
21. BS(2014-1) also present an alternative version of their model in which both households and experts have logarithmic preferences (once again this choice of preferences simplifies solution because the value function is then additively separable and optimal consumption is a fixed proportion of the market value of agent net worth). As long as experts are more impatient than households this generates very similar dynamics to the baseline model, but now with positive expert consumption (i.e., some payment of dividends) for all values of the state variable.
22. See our Section 4.2.2 and Appendix A.3 for discussion of the calculation of this ergodic density.
23. Phelan (2015) also introduces the banking sector in a continuous-time macrofinancial model, however, without explicitly modeling this lending channel.
24. Both loans and deposits are assumed to be short term, and the full depreciation of productive capital is allowed.
25. The distribution of equity capital across individual banks then has no impact on economy wide outcomes.
26. It is noteworthy that many of the BS(2014-2) results were originally obtained using a quite different underlying model of risks to bank asset returns, based on Poisson shocks, see Brunnermeier and Sannikov (2014c).
27. Of course, any invertible function of x could be considered the macrostate as well. In this particular example, one could alternatively use the capital price q as a state variable, since the mapping between q and x is invertible.
28. Brunnermeier and Sannikov (2014a) obtain the same boundary condition as follows: At $x = 0$ experts get excess returns of $a/q(0) - r > 0$. Choosing φ high enough, their rate of returns exceeds their discount rate ρ, and value function becomes infinite. However, since x can never escape from 0, and experts only consume at $x = x^*$, it is not totally clear that V can indeed grow unboundedly. The condition is therefore plausible but may require more careful analysis to be rigorously justified.
29. In a discrete time set-up an attempt to accommodate this feature is made by Rampini and Viswanathan (2012).
30. The above uses the Itô interpretation of the SDE (A.32) which amounts to assuming that the noise amplitude σ is evaluated at the start of each period. In general one can set $x_{(k+1)T} = x_{kT} + T\mu(x_{kT}) + \sqrt{T}\sigma(x_{kT+\alpha})\eta_{kT}$ where $0 \leq \alpha \leq 1$. Choice of α does influence the form of later formulas. The Itô interpretation, $\alpha = 0$ is the default choice in economics applications, as equations of motion are supposed not to pre-empt the shocks. In natural sciences, the Stratonovich convention $\alpha = 1/2$ is commonly used.
31. A standard approach is recasting the SDE with reflection into a so called Skorokhod problem, whereby the process x is seen as driven by an additional process k, $\mathrm{d}x_t = \mu(x_t)\,\mathrm{d}t + \sigma(x_t)\,\mathrm{d}z_t + \mathrm{d}k_t$, and where $\mathrm{d}k_t$ is non-zero only on the boundary. See e.g., Lions and Sznitman (1984).

References

Tobias Adrian and Nina Boyarchenko. Interdiary Leverage Cycles and Financial Stability. 2013. URL http://www.newyorkfed.org/research/staff_reports/sr567.pdf.

Philippe Aghion and Peter Howitt. A Model of Growth Through Creative Destruction. *Econometrica*, 60(2):323–351, March 1992. ISSN 0012-9682. doi: 10.2307/2951599. URL http://www.jstor.org/stable/2951599.

Philippe Aghion and Peter Howitt. *The Economics of Growth*. MIT Press, Boston, 2009.

S Rao Aiyagari. Uninsured idiosyncratic risk and aggregate saving. *The Quarterly Journal of Economics*, 109(3):659–684, 1994. ISSN 0033-5533.

Yann Algan, Olivier Allais, Wouter J Den Haan, and Pontus Rendahl. Solving and simulating models with heterogeneous agents and aggregate uncertainty. *Handbook of Computational Economics*, 2010.

Bo Becker and Victoria Ivashina. Cyclicality of credit supply: Firm level evidence. *Journal of Monetary Economics*, 62:76–93, March 2014. ISSN 0304-3932. doi: 10.1016/j.jmoneco.2013.10.002.

Ben Bernanke, Mark Gertler, and Simon Gilchrist. The Financial Accelerator in a Quantitative Business Cycle Framework. In John B Taylor and Michael Woodford, editors, *Handbook of Macroeconomics, Volume 1C*, pages 1341–1393. Elsevier Science, North-Holland, 1999.

Fischer Black and Myron Scholes. The pricing of options and corporate liabilities. *The Journal of Political Economy*, pages 637–654, 1973. ISSN 0022-3808.

Patrick Bolton, Hui Chen, and Neng Wang. A Unified Theory of Tobin's q, Corporate Investment, Financing, and Risk Management. *The Journal of Finance*, 66(5): 1545–1578, October 2011. ISSN 0022-1082. doi: 10.1111/j.1540–6261.2011.01681.x. URL http://doi.wiley.com/10.1111/j.1540-6261.2011.01681.x.

Markus K Brunnermeier and Yuliy Sannikov. Macro, Money and Finance: A Continuous-Time Approach. In John Taylor and Harald Uhlig, editors, *Handbook of Macroeconomics Vol 2* 2016 forthcoming.

Markus K. Brunnermeier and Yuliy Sannikov. International Credit Flows and Pecuniary Externalities. *American Economic Journal: Macroeconimics*, 7(1): 297–338, 2015

Markus K. Brunnermeier and Yuliy Sannikov. A Macroeconomic Model with a Financial Sector. *American Economic Review*, 104(2):379–421, February 2014a. ISSN 0002-8282. doi: 10.1257/aer.104.2.379. URL http://www.ingentaconnect.com/content/aea/aer/2014/00000104/00000002/art00002.

Markus K Brunnermeier and Yuliy Sannikov. A Macroeconomic Model with a Financial Sector: online appendix. *American Economic Review*, (2), February 2014b. ISSN 0002-8282. doi: 10.1257/aer.104.2.379. URL http://www.ingentaconnect.com/content/aea/aer/2014/00000104/00000002/art00002.

Markus K Brunnermeier and Yuliy Sannikov. The I Theory of Money: version of April 2014. 2014c.

Markus K Brunnermeier and Yuliy Sannikov. The I Theory of Money: version of November 2014. 2014d.

Markus K Brunnermeier, Thomas M Eisenbach, and Yuliy Sannikov. Macroeconomics with financial frictions: A survey. Technical report, 2012.

Guillermo A. Calvo. Capital Market Crises. *Journal of Applied Economics*, 1(1):35–54, 1998.

George M Constantinides. Intertemporal asset pricing with heterogeneous consumers and without demand aggregation. *Journal of Business*, pages 253–267, 1982. ISSN 0021-9398.

Michael G Crandall, Ishii Hitoshi, and P.-L. Lions. User's guide to viscosity solutions of second order partial differential equations. *Bulletin of the American Mathematical Society*. 27:1–67, 1992.

Harry DeAngelo and René M Stulz. Why high leverage is optimal for banks. Technical report, 2013.

Wouter J. Den Haan. Comparison of solutions to the incomplete markets model with aggregate uncertainty. *Journal of Economic Dynamics and Control*, 34(1):4–27, January 2010. ISSN 0165-1889. doi: 10.1016/j.jedc.2008.12.010. URL http://www.sciencedirect.com/science/article/pii/S0165188909001298.

Douglas Diamond and Phillip Dybvig. Bank runs, deposit insurance and liquidity. *Journal of Political Economy*, 91:401–419, 1983.

Douglas W. Diamond. Financial Intermediation and Delegated Monitoring. *The Review of Economic Studies*, 51(3):393, July 1984. ISSN 0034-6527. doi: 10.2307/2297430. URL http://restud.oxfordjournals.org/content/51/3/393.short.

William Feller. *An Introduction to Probability Theory and its Applications*, volume II. Wiley, New York, 2nd edition, 1971.

Zhigang Feng, Jianjun Miao, Adrian Peralta-Alva, and Manuel S Santos. Numerical simulation of nonoptimal dynamic equilibrium models. *International Economic Review*, 55(1):83–110, February 2014. ISSN 1468-2354. doi: 10.1111/iere.12042. URL http://dx.doi.org/10.1111/iere.12042.

Wendell Fleming and Halil Mete Soner. *Controlled {Markov} Processes and Viscosity Solutions*. Stochastic Modelling and Applied Probability. Springer, New York, 2nd edition, 2006.

Marvin Frankel. The production function in allocation and growth: a synthesis. *The American Economic Review*, pages 996–1022, 1962. ISSN 0002-8282.

Crispin Gardiner. *Stochastic Methods: A Handbook for the Natural and Social Sciences*. Springer Series in Synergetics. Springer, Berlin, 4th edition, 2009.

William M Gorman. Separable utility and aggregation. *Econometrica: Journal of the Econometric Society*, pages 469–481, 1959. ISSN 0012-9682.

Luca Guerrieri and Matteo Iacoviello. OccBin: A toolkit for solving dynamic models with occasionally binding constraints easily. *Journal of Monetary Economics*, 70: 22–38, March 2015. ISSN 0304-3932. doi: 10.1016/j.jmoneco.2014.08.005. URL http://www.sciencedirect.com/science/article/pii/S0304393214001238.

Fatih Guvenen. Macroeconomics with hetereogeneity: a practical guide. *Economic Quarterly*, (3Q):255–326, 2011.

Ernst Hairer, Gerhard Wanner, and Syvert P Nø rsett. *Solving Ordinary Differential Equations I: Nonstiff Problems*. Springer Series in Computational Mathematics. Springer, Berlin, 1993.

Andrew G Haldane. On microscopes and telescopes, 2015. URL http://www.bankofengland.co.uk/publications/Documents/speeches/2015/speech812.pdf.

Zhigu He and Arvind Krishnamurthy. A model of capital and crises. *The Review of Economic Studies*, 79(2):735–777, 2012. ISSN 0034-6527.

Zhiguo He and Arvind Krishnamurthy. Intermediary asset pricing. *American Economic Review*, 103(2):732–770, 2013. ISSN 0002-8282.

Zhiguo He and Arvind Krishnamurthy. A macroeconomic framework for quantifying systemic risk. 2014, NBER Working Paper No. 19885, February.

B Holmstrom and J Tirole. Financial intermediation, loanable funds, and the real sector. *Quarterly Journal of Economics*, 112(3):663–691, 1997.

Mark Huggett. The risk-free rate in heterogeneous-agent incomplete-insurance economies. *Journal of Economic Dynamics and Control*, 17(5):953–969, 1993. ISSN 0165-1889.

Jukka Isohätälä, Feodor Kusmartsev, Alistair Milne, and Donald Robertson. Leverage constraints and real interest rates. *The Manchester School* 2015. URL http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2526790.

Jukka Isohätälä, Alistair Milne, and Donald Robertson. The Net Worth Trap: Investment and Output Dynamics in the Presence of Financing Constraints. 2014. URL http://www.suomenpankki.fi/en/julkaisut/tutkimukset/keskustelualoitteet/Pages/dp2014_26.aspx.

Nobuhiro Kiyotaki and John Moore. Credit cycles. *Journal of Political Economy*, 105(2): 211–248, 1997.

Fima C Klebaner. *Introduction to Stochastic Calculus with Applications*. Imperial College Press, 2nd edition, 2005, Working Paper, University of Zurich.

Nataliya Klimenko, Sebastian Pfeil, and Jean-Charles Rochet. Bank Capital and Aggregate Credit. 2015.

Dirk Krueger and Felix Kubler. Markov equilibria in macroeconomics. In *The New Palgrave Dictionary of Economics'*, Palgrave Macmillan, Basingstoke. 2008.

Per Krusell and Anthony A Smith. Income and wealth heterogeneity, portfolio choice, and equilibrium asset returns. *Macroeconomic dynamics*, 1(02):387–422, 1997. ISSN 1469-8056.

Per Krusell and Anthony A Jr Smith. Income and wealth heterogeneity in the macroeconomy. *Journal of Political Economy*, 106(5):867–896, 1998. ISSN 0022-3808.

Ricardo Lagos and Randall Wright. A Unified Framework for Monetary Theory and Policy Analysis. *Journal of Political Economy*, 113(3):463–484, 2005.

P L Lions and A S Sznitman. Stochastic differential equations with reflecting boundary conditions. *Communications on Pure and Applied Mathematics*, 37(4):511–537, 1984. ISSN 1097-0312. doi: 10.1002/cpa.3160370408.

Lars Ljungqvist and Thomas J Sargent. *Recursive Macroeconomic Theory*. MIT Press, 2nd edition, 2000.

Robert E Lucas and Nancy L Stokey. Money and Interest in a Cash-in-Advance Economy. *Econometrica*, 55(3):491–513, 1987.

Robert E Lucas Jr. Asset prices in an exchange economy. *Econometrica*, 46(6): 1429–1445, November 1978. ISSN 0012-9682. doi: 10.2307/1913837. URL http://www.jstor.org/stable/1913837.

N Gregory Mankiw. The equity premium and the concentration of aggregate shocks. *Journal of Financial Economics*, 17(1):211–219, 1986. ISSN 0304-405X.

Césaire A Meh and Kevin Moran. The Role of Bank Capital in the Propagation of Shocks. *Journal of Economic Dynamics and Control*, 34(3):555–576, 2010.

Robert C Merton. Lifetime Portfolio Selection under Uncertainty: The Continuous-Time Case. *Review of Economics and Statistics.*, 51(3):247–257, 1969. doi: 10.2307/1926560.

Robert C. Merton. Optimum consumption and portfolio rules in a continuous-time model. *Journal of Economic Theory*, 3(4):373–413, 1971. URL http://ideas.repec.org/a/eee/jetheo/v3y1971i4p373-413.html.

Bernt Øksendal. *Stochastic Differential Equations: An Introduction with Applications*. Universitext. Springer, Berlin, 6th edition, 2003.

Gregory Phelan. Financial intermediation, leverage, and macroeconomic instability. *American Economic Journal: Macroeconomics*, forthcoming, 2015.

Adriano A Rampini and S Viswanathan. Financial intermediary capital. *Available at SSRN 1785877*, 2012.

Carmen Reinhart and Kenneth Rogoff. *This Time Is Different: Eight Centuries of Financial Folly*. Princeton University Press, 2009. URL http://www.amazon.co.uk/This-Time-Different-Centuries-Financial-ebook/dp/B004EYT932/ref=sr_1_1?ie=UTF8&qid=1404847123&sr=8-1&keywords=this+time+is+different+rogoff.

Hannes Risken. *The {Fokker-Planck} Equation: Methods of Solution and Applications*. Springer Series in Synergetics. Springer, Berlin, 1996.

Paul M Romer. Growth Based on Increasing Returns Due to Specialization. *American Economic Review*, 77(2):56–62, 1987. ISSN 0002-8282.

Paul M Romer. Endogenous Technological Change. *Journal of Political Economy*, 98(5): S71–102, 1990. ISSN 0022-3808.

Mark Rubinstein. An aggregation theorem for securities markets. *Journal of Financial Economics*, 1(3):225–244, 1974. ISSN 0304-405X.

Paul A Samuelson. An exact consumption-loan model of interest with or without the social contrivance of money. *The Journal of Political Economy*, pages 467–482, 1958. ISSN 0022-3808.

F Smets and R Wouters. Comparing Shocks and Frictions in {US} and Euro Area Business Cycles: A {B}ayesian {DSGE} Approach. *Journal of Applied Econometrics*, 20:161–183, 2005.

M Woodford. *Interest and Prices: Foundations of a Theory of Monetary Policy*. Princeton University Press, 2003.

11
Recent Results on Operator Techniques in the Description of Macroscopic Systems

Fabio Bagarello

1 Introduction

In many classical systems the relevant quantities we are interested in change discontinuously. For instance, if you consider a certain *population* $\mathcal{P}$, and its time evolution, the number of *people* forming $\mathcal{P}$ cannot change arbitrarily: if, at $t_0 = 0$, $\mathcal{P}$ consists of N_0 elements, at some later time $t_1 > t_0$, $\mathcal{P}$ may only consist of N_1 elements, with N_1 differing from N_0 for an integer quantity. The same happens if our system consists of two (or more) different populations, $\mathcal{P}_1$ and $\mathcal{P}_2$ (e.g., preys and predators or two migrating species): again, the total number of their elements can only take, for obvious reasons, integer values.

Something similar happens for a *simplified stock market* (SSM), which, for us, is a group of people (the *traders*) with some money and a certain number of shares of different kinds, which are exchanged between the traders. Of course, they pay some cash for that. Also in this case it is clear that natural numbers play a crucial role: in the SSM a trader may have only a natural number of shares (30, 5000 or 10^6, but not .75 shares), and a natural number of *units of cash* (there is nothing less than one cent of euro, for instance). Hence, if two traders buy or sell a share, the number of shares in their portfolios increases or decreases by one unit, and the amount of their money changes for an integer multiple of the unit of cash. This might appear (but is not!) just a simple discretization of a continuous problem, for which several approaches have been proposed along the years. In fact, we adopt here a rather different philosophy, which can be summarized as follows: the discrete quantities used in the description of the system $\mathcal{S}$ under analysis are closely related to the eigenvalues of some (mainly) self-adjoint operator. Of course, a crucial and natural question is how the dynamical behavior of $\mathcal{S}$ can be deduced. Along all our work we have chosen to use an Heisenberg-like dynamics, or its Schrödinger counterpart, which we believe is the good choice because of the following reasons:

1. It is the choice for quantum mechanical systems, where operators are essential.
2. It is usually quite easy to write down an energy-like operator, Hamiltonian H of the system $\mathcal{S}$, which determines the dynamics of the system. This is, in fact, the content of Section 2.1. Among the other criteria, the explicit definition of H is suggested by the existence of some conserved quantities of $\mathcal{S}$: if X is an operator which is expected to be preserved during the time evolution of $\mathcal{S}$, like for instance the total amount of cash in a closed SSM, then, because of the definition of the Heisenberg dynamics, H must commute with X: $[H,X]=0$. This gives some extra hints on how to define H explicitly. Another criterion is the following: H must *contains* in itself the main phenomena we have to describe. This is what is usually done in many body theory or in elementary particle physics, where, e.g., the possibility of two particles to interact is reflected in the presence in H of a term which describe this interaction [1]. Once H has been determined, it can be used to find the time evolution of any *observable* A of $\mathcal{S}$ using the standard Heisenberg prescription: $A(t)=e^{iHt}A(0)e^{-iHt}$, $A(0)$ being the value of A at $t=0$.
3. It produces results which, at least for some *easy* systems, look quite reasonable, since they are observed in real life, see Section 3.

It is worth stressing that, since all the observables we are usually interested in in our applications form a *commuting* subset of a larger non-abelian algebra, they can be diagonalized simultaneously. Therefore, for our purposes, the eigenstates of these commuting observables form an orthonormal basis of the Hilbert space $\mathcal{H}$ used in the description of $\mathcal{S}$. This commutativity implies that, in the complete description of $\mathcal{S}$, all the results which are deduced using our approach are not affected by any uncertainty, as one could possibly expect. In other words, while not commuting observables imply some Heisenberg-like uncertainty relation, this is not the case for us, due to the fact that all our observables do commute.

Remark:– In some specific applications, the impossibility of observing simultaneously two (apparently) classical quantities has been taken as a strong indication of the relevance of a quantum-like structure in the description of that process, showing, in particular, the importance of non-commuting operators. This is what was proposed, for instance, in [2], where the authors assume that a trader in a realistic market cannot be able to know, at the same time, the price of a certain share and its forward time derivative. The reason is clear: if the trader has access to both these information with absolute precision, then he is surely able to earn as much as he wants! For this reason, W. Segal and I. E. Segal proposed to use two non-commuting operators to describe the price and its time derivative.

It is surely interesting to observe that, in the last few years, a growing interest in classical applications of quantum ideas appeared in the literature. Some recent monographs along this line are [3, 4, 5, 6, 7].

This chapter is organized as follows: in the next section we introduce those quantum tools which are useful in our framework, focusing in particular on the canonical commutation rules and on the quantum dynamics. Moreover, we will also discuss in many details how the Hamiltonian of a given system should be constructed, considering both the case of closed and open systems. Our framework is then applied to two completely different situations, i.e., to the description of a love story, which is a first simple (and not-so-simple) dynamical model, and to the analysis of a SSM, which is, not surprisingly, much harder to treat. Section 6 contains our conclusions.

2 Our quantum tools

This section is dedicated to introducing the tools and the rules we are going to use in the rest of this paper: Let $\mathcal{S}$ be our physical system and let $\mathfrak{A}$ the set of all the operators useful for a complete description of $\mathcal{S}$, which includes the observables of $\mathcal{S}$, i.e., those quantities which are measured in a concrete experiment. Let $\mathcal{H}$ be the Hilbert space of the system. It might happen, and often happens, that if X is an observable of $\mathcal{S}$, X is unbounded.

As already said, a particularly relevant role in our strategy is played by a suitable self-adjoint operator $H = H^\dagger$ *attached* to $\mathcal{S}$, called *the Hamiltonian* of $\mathcal{S}$. In standard quantum mechanics, H represents the energy of $\mathcal{S}$. In most cases, H is unbounded. In the so-called *Heisenberg representation*, the time evolution of an observable $X \in \mathfrak{A}$ is given by

$$X(t) = e^{iHt}Xe^{-iHt} \tag{11.1}$$

or, equivalently, by the solution of the differential equation

$$\frac{dX(t)}{dt} = ie^{iHt}[H,X]e^{-iHt} = i[H,X(t)], \tag{11.2}$$

where $[A,B] := AB - BA$ is the *commutator* between A and B. Notice that $e^{\pm iHt}$ are unitary operators, hence they are bounded. The time evolution defined in this way is usually a one parameter group of automorphisms of $\mathfrak{A}$: for each $X,Y \in \mathfrak{A}$, and for all $t,t_1,t_2 \in \mathbb{R}$, $(XY)(t) = X(t)Y(t)$ and $X(t_1 + t_2) = (X(t_1))(t_2)$. An operator $Z \in \mathfrak{A}$ is a *constant of motion* if it commutes with H. Indeed, in this case, Equation (11.2) implies that $\dot{Z}(t) = 0$, so that $Z(t) = Z(0)$ for all t. It is worth stressing that, in formulas (11.1) and (11.2), we are assuming that H does not depend explicitly on time, which is not always true.

A very special role in our framework is played by the so-called *canonical commutation relations* (CCRs): we say that a set of operators $\{a_l, a_l^\dagger, l = 1,2,\dots,L\}$, acting on the Hilbert space $\mathcal{H}$, satisfy the CCRs, if the following hold:

$$[a_l, a_n^\dagger] = \delta_{ln} 1\!\!1, \qquad [a_l, a_n] = [a_l^\dagger, a_n^\dagger] = 0, \tag{11.3}$$

for all $l, n = 1, 2, \ldots, L$, $1\!\!1$ being the identity operator on $\mathcal{H}$. These operators are those which are used to describe L different *modes* of bosons. From these operators we can construct $\hat{n}_l = a_l^\dagger a_l$ and $\hat{N} = \sum_{l=1}^{L} \hat{n}_l$ which are both self-adjoint. In particular $\hat{n}_l$ is the *number operator* for the l-th mode, while $\hat{N}$ is the *number operator* for $\mathcal{S}$. The reason for this terminology will appear clear later on.

An orthonormal (o.n.) basis of $\mathcal{H}$ can be constructed as follows: we introduce the *vacuum* of the theory, that is a vector $\varphi_{\mathbf{0}}$ which is annihilated by all the operators a_l: $a_l \varphi_{\mathbf{0}} = 0$ for all $l = 1, 2, \ldots, L$. Then we act on $\varphi_{\mathbf{0}}$ with the operators $a_l^\dagger$ and with their powers,

$$\varphi_{n_1, n_2, \ldots, n_L} := \frac{1}{\sqrt{n_1! n_2! \ldots n_L!}} (a_1^\dagger)^{n_1} (a_2^\dagger)^{n_2} \cdots (a_L^\dagger)^{n_L} \varphi_{\mathbf{0}}, \tag{11.4}$$

$n_l = 0, 1, 2, \ldots$, for all l, and we normalize the vectors obtained in this way. The set of the $\varphi_{n_1, n_2, \ldots, n_L}$'s forms a complete and o.n. set in $\mathcal{H}$, and they are eigenstates of both $\hat{n}_l$ and $\hat{N}$:

$$\hat{n}_l \varphi_{n_1, n_2, \ldots, n_L} = n_l \varphi_{n_1, n_2, \ldots, n_L}$$

and

$$\hat{N} \varphi_{n_1, n_2, \ldots, n_L} = N \varphi_{n_1, n_2, \ldots, n_L},$$

where $N = \sum_{l=1}^{L} n_l$. Hence, n_l and N are eigenvalues of $\hat{n}_l$ and $\hat{N}$ respectively. Moreover, using the CCRs we deduce that

$$\hat{n}_l \left(a_l \varphi_{n_1, n_2, \ldots, n_L} \right) = (n_l - 1)(a_l \varphi_{n_1, n_2, \ldots, n_L}),$$

for $n_l \geq 1$ while, if $n_l = 0$, a_l annihilates the vector, and

$$\hat{n}_l \left(a_l^\dagger \varphi_{n_1, n_2, \ldots, n_L} \right) = (n_l + 1)(a_l^\dagger \varphi_{n_1, n_2, \ldots, n_L}),$$

for all l and for all n_l. For these reasons the following interpretation is given in the literature: if the L different modes of bosons of $\mathcal{S}$ are described by the vector $\varphi_{n_1, n_2, \ldots, n_L}$, this means that n_1 bosons are in the first mode, n_2 in the second mode, and so on. The operator $\hat{n}_l$ acts on $\varphi_{n_1, n_2, \ldots, n_L}$ and returns n_l, which is exactly the number of bosons in the l-th mode. The operator $\hat{N}$ counts the total number of bosons. Moreover, the operator a_l destroys a boson in the l-th mode, while $a_l^\dagger$ creates a boson in the same mode: a_l and $a_l^\dagger$ are called the *annihilation* and the *creation* operators.

The vector $\varphi_{n_1, n_2, \ldots, n_L}$ in (11.4) defines a *vector (or number) state* over the set $\mathfrak{A}$ as

$$\omega_{n_1, n_2, \ldots, n_L}(X) = \langle \varphi_{n_1, n_2, \ldots, n_L}, X \varphi_{n_1, n_2, \ldots, n_L} \rangle, \tag{11.5}$$

where $\langle , \rangle$ is the scalar product in the Hilbert space $\mathcal{H}$. These states will be used to *project* from quantum to classical dynamics and to fix the initial conditions of the system under consideration, in a way which will be clarified later on.

Notice that a_l, $a_l^\dagger$, $\hat{n}_l$ and $\hat{N}$, are all unbounded, and therefore, they have severe domain problems, since they cannot be defined in all of $\mathcal{H}$. However, each vector $\varphi_{n_1,n_2,\ldots,n_L}$ belongs to the domains of all the operators which are relevant for us. Moreover, in some applications $\mathcal{H}$ can be replaced by an *effective* Hilbert space, $\mathcal{H}_{eff}$, which becomes *dynamically* finite-dimensional because of the existence of some conserved quantities and because of the initial conditions, which imposes some constraints on the accessible (energy) levels [8].

Remark:– In some applications, rather than CCRs, it is convenient to use the so-called *canonical anti-commutation relations* (CARs), which are defined by a set of operators $\{b_l, b_l^\dagger, \ell = 1,2,\ldots,L\}$, acting on a certain finite-dimensional Hilbert space $\mathcal{H}_F$ and satisfying the following rules:

$$\{b_l, b_n^\dagger\} = \delta_{l,n} 1\!\!1, \qquad \{b_l, b_n\} = \{b_l^\dagger, b_n^\dagger\} = 0,$$

for all $l,n = 1,2,\ldots,L$. Here, $\{x,y\} := xy + yx$ is the *anticommutator* of x and y and $1\!\!1$ is now the identity operator on $\mathcal{H}_F$. However, in the applications considered in this chapter we will not need to use CARs, so we refer to [6] for more details and applications.

2.1 Writing the Hamiltonian

Apart from the general functional settings of our framework, the main ingredient for the dynamical description of the system $\mathcal{S}$ is surely its Hamiltonian H. Now, following [6], we describe in some details the scheme we use to write down H in different contexts, starting from closed and moving to open systems.

2.1.1 The Hamiltonian for closed systems

Our closed system $\mathcal{S}$ is made of *elements* τ_j, $j = 1,2,\ldots,N_{\mathcal{S}}$, which are only interacting among themselves. Each element τ_j is defined by a certain set of variables (the self-adjoint *observable operators* for τ_j): $(v_1^{(j)}, v_2^{(j)}, \ldots, v_{K_j}^{(j)})$. In other words, the K_j-dimensional operator-valued vector $\mathbf{v}^{(j)} := (v_1^{(j)}, v_2^{(j)}, \ldots, v_{K_j}^{(j)})$ defines completely the dynamical status of τ_j.

Example 1:– In Section 3 we are interested in a love triangle in which Bob interacts with Alice and with Carla. In this case τ_1 is Bob, τ_2 is Alice and τ_3 is Carla. Then $v_1^{(1)}$ is Bob's *level of attraction* (LoA) for Alice, $v_1^{(1)} = \hat{n}_{12} = a_{12}^\dagger a_{12}$, while $v_2^{(1)}$ is his LoA for Carla, $v_2^{(1)} = \hat{n}_{13} = a_{13}^\dagger a_{13}$. We will be more precise about their meaning later on. Here we just want to say that a_{1j}, $j = 1,2$, are bosonic operators as described before. Then Bob's status is completely described by a two-dimensional vector (here $K_1 = 2$), whose components are the two LoA's which Bob experiences for Alice and Carla: nothing more is required to describe

Bob, and for this reason his dynamical behavior is entirely known when the time evolution of the vector $\mathbf{v}^{(1)} := (v_1^{(1)}, v_2^{(1)}) = (\hat{n}_{12}, \hat{n}_{13})$ is obtained. As for Alice, her status is described by a one dimensional vector, $\mathbf{v}^{(2)} := (v_1^{(2)}) = (\hat{n}_2)$, whose single component is simply Alice's LoA for Bob. Then $K_2 = 1$. Similarly, for Carla, $K_3 = 1$, and her status is again another one dimensional vector, $\mathbf{v}^{(3)} := (v_1^{(3)}) = (\hat{n}_3)$, whose unique component is Carla's LoA for Bob.

Example 2:– In Section 4 we describe a SSM, made by $N_{\mathcal{S}}$ traders τ_j, $j = 1, 2, \ldots, N_{\mathcal{S}}$, and the status of each one of these traders is defined by the number of shares and by the amount of money in their *portfolios*, see below. If, for simplicity, we assume that just a single type of shares goes around the SSM, then the vector $\mathbf{v}^{(j)}$ defining the status of τ_j is the following

$$\mathbf{v}^{(j)} = \left(\hat{n}_j, \hat{k}_j\right),$$

where $\hat{n}_j$ and $\hat{k}_j$ are the shares and the cash operators for the portfolio of τ_j. Their mean values will be used to compute the explicit value of this portfolio. As in the previous example, see Section 4, $\hat{n}_j$ and $\hat{k}_j$ can be written in terms of bosonic operators.

The time evolution of $\mathcal{S}$ follows from the time evolution of (each) vector $\mathbf{v}^{(j)}$, which can be deduced by a suitable Hamiltonian, whose analytical expression is fixed by considering some guiding rules, which we discuss below.

The first natural requirement is our rule

R1:– *in absence of interactions between the different τ_j's, all the vectors $\mathbf{v}^{(j)}$ stay constant in time.*

Stated in different words, Rule **R1** means that, if there is no interaction Hamiltonian, then the free Hamiltonian H_0 must commute with each component of the various $\mathbf{v}^{(j)}$, $v_i^{(j)}, j = 1, 2, \ldots, N_{\mathcal{S}}$, $i = 1, 2, \ldots, K_j$. For instance, considering the love triangle in Example 1, we should have

$$[H_0, \hat{n}_{12}] = [H_0, \hat{n}_{13}] = [H_0, \hat{n}_2] = [H_0, \hat{n}_3] = 0.$$

This is reasonable since, if there is no interaction between Alice, Bob and Carla, there is no reason for their LoA's to change with time. This is a very general and natural rule for us: only the interactions cause a significant change in the status of the system.

Let us now consider $\mathcal{S}$ *globally*, i.e., looking at the set of all the τ_j's. Depending on the system we are considering, it might happen that, for some reason, some global function of the $\mathbf{v}^{(j)}$'s is expected to stay constant in time. For instance, when dealing with a closed SSM, two such functions surely exist: the total number of shares and the total amount of cash. In fact, if the market is closed, both the cash and the shares can only be exchanged between the traders, while they cannot be created or destroyed. Considering the Alice-Bob's (closed) love affair, the idea we will adopt in the construction of the model is that the sum of Bob's and Alice's LoA's stay constant in time. This is a simple way to state our *law of attraction*, introduced in Section 3. These considerations are behind our second rule:

R2:– *if a certain global function $f(\mathbf{v}^{(1)},\mathbf{v}^{(2)},\ldots,\mathbf{v}^{(N_{\mathcal{S}})})$ is expected to stay constant in time, then H must commute with f: $\left[H,f(\mathbf{v}^{(1)},\ldots,\mathbf{v}^{(N_{\mathcal{S}})})\right]=0$.*

Notice that here, unlike to what we have done in **R1**, here H is the full Hamiltonian and not just is free part H_0: in fact, the interactions are the interesting part of the problem. For instance, in the simple love affair described in Section 3, the Hamiltonian of the interacting system, H, should commute with $a_1^\dagger a_1 + a_2^\dagger a_2$, which represents the *global LoA*. Analogously, if c_j and a_j are the annihilation operators associated to the cash and to the shares in the portfolio of the trader τ_j, see Section 4, then, calling $\hat{K}=\sum_{j=1}^{N_{\mathcal{S}}} c_j^\dagger c_j$ and $\hat{N}=\sum_{j=1}^{N_{\mathcal{S}}} a_j^\dagger a_j$ the *global cash and shares operators*, H must commute with both: $[H,\hat{K}]=[H,\hat{N}]=0$.

Suppose now that τ_1 and τ_2 interact and that, because of this interaction, they change their status. For instance, the initial vector $<\mathbf{v}^{(1)}(b.i.)>$, after the interaction is replaced by $<\mathbf{v}^{(1)}(a.i.)>$. Here *b.i.* and *a.i.* stand respectively for *before* and *after interaction*, and $<\mathbf{v}>$ means that we are considering the mean value of the operator-valued vector $\mathbf{v}$ on a suitable state, see (11.5) for instance. Analogously, $<\mathbf{v}^{(2)}(b.i.)>$ changes to $<\mathbf{v}^{(2)}(a.i.)>$. To be concrete, we suppose here that the vectors $<\mathbf{v}^{(1)}(a.i.)>$ and $<\mathbf{v}^{(1)}(b.i.)>$ differ only for the values of their first two components, those related to τ_1 and τ_2, $<v_1^{(1)}>$ and $<v_2^{(1)}>$. We call $\delta v_j^{(1)} =< v_j^{(1)}(a.i.)> - < v_j^{(1)}(b.i)>$, $j=1,2$, the differences between these values. We introduce now $a_1^{(1)}$ and $a_2^{(1)}$, two (bosonic or fermionic) annihilation operators associated to these components. Analogously, let us suppose that $<\mathbf{v}^{(2)}(a.i.)>$ and $<\mathbf{v}^{(2)}(b.i.)>$ also differ only for the values of their first two components, $<v_1^{(2)}>$ and $<v_2^{(2)}>$, and let $a_1^{(2)}$ and $a_2^{(2)}$ be the related annihilation operators, and let $\delta v_j^{(2)} =< v_j^{(2)}(a.i.)> - < v_j^{(2)}(b.i)>$, $j=1,2$. To build up the interaction Hamiltonian H_{int} which is responsible for such a change, we first need to consider the signs of the various $\delta v_j^{(k)}$. To fix the ideas, we suppose here that $\delta v_1^{(1)}$ and $\delta v_2^{(2)}$ are positive, while $\delta v_2^{(1)}$ and $\delta v_1^{(2)}$ are negative. Then our third rule can be stated as follows:

R3:– *the interaction Hamiltonian H_{int} responsible for the above changes must contain the contribution*

$$\left(a_1^{(1)\dagger}\right)^{\delta v_1^{(1)}} \left(a_2^{(1)}\right)^{\delta v_2^{(1)}} \left(a_1^{(2)}\right)^{\delta v_1^{(2)}} \left(a_2^{(2)\dagger}\right)^{\delta v_2^{(2)}}, \tag{11.6}$$

together with its Hermitian conjugate.

Introducing the Hermitian conjugate in our Hamiltonians is the easiest way to get an unitary evolution, and this is crucial if we want to use the Heisenberg equation of motion (11.2). Looking at (11.6) we observe that creation operators appears for positive values of $\delta v_j^{(k)}$'s, while annihilation operators are used for negative $\delta v_j^{(k)}$'s. The reason is clear: since, for instance, $\delta v_1^{(1)}$ is positive, during the interaction between τ_1 and τ_2 the value of $< v_1^{(1)} >$ increases, and this increment can only be produced (in our settings) by creation operators. On the other hand, since $\delta v_1^{(2)}$ is negative, the value of $< v_1^{(2)} >$ decreases, and this is well described by considering an annihilation operator in H_{int}. Of course, it is not particularly difficult to extend (11.6) to slightly different situations, for instance when more components of the status vectors are involved in the interaction between τ_1 and τ_2. It is also very simple to go from this, in general, non-linear interaction Hamiltonian to its linear version: this is what happens when all the variation parameters $\delta v_j^{(k)}$ are equal to one. From an analytical point of view, this is often a particularly simple situation since, most of the times, the computations can be carried out analytically without the need of any approximation and/or numerical techniques.

The Hamiltonian in (11.8) below is a first example of how Rule **R3** is implemented in a simple situation, where the non linearity is *related* to a single variable and to a single actor, Bob. More complicated examples, see (11.10) or (11.14), will also be introduced later. Of course, the difficulties of the dynamics deduced by these Hamiltonians will be directly proportional to the complexity of the system under consideration.

As we have already discussed several times so far, looking at a certain classical system $\mathcal{S}$ as if it was a closed system, it might not necessarily be the best point of view. In fact, quite often, the role of the environment turns out to be crucial in producing a reasonable dynamical behavior for $\mathcal{S}$. For instance, open systems may reach some equilibrium, while (finite-dimensional) closed systems always oscillate.

As in the standard literature, an open system for us is simply a system $\mathcal{S}$ interacting with a reservoir $\mathcal{R}$. From a dynamical point of view, the Hamiltonian $H_{\tilde{\mathcal{S}}}$ of this larger system, $\tilde{\mathcal{S}} := \mathcal{S} \cup \mathcal{R}$, appears to be the Hamiltonian of an interacting system. This means that the general expression of $H_{\tilde{\mathcal{S}}}$ is the sum of three contributions:

$$H_{\tilde{\mathcal{S}}} = H_{\mathcal{S}} + H_{0,\mathcal{R}} + H_{\mathcal{S},\mathcal{R}}. \tag{11.7}$$

Here $H_{\mathcal{S}}$ is the Hamiltonian for $\mathcal{S}$, whose explicit expression should be deduced adopting rules **R1**, **R2** and **R3**, working as if $\mathcal{S}$ was a closed system by itself. This implies, among other things, that $H_{\mathcal{S}}$ contains a free contribution plus a second term describing the interactions among the various elements of $\mathcal{S}$. $H_{0,\mathcal{R}}$ is the free Hamiltonian of the reservoir, while $H_{\mathcal{S},\mathcal{R}}$ contains the interactions between the system and the reservoir. To fix the explicit expression of $H_{0,\mathcal{R}}$ we adopt the following rule, which extends to the reservoir our previous rule **R1**:

R4:– *in the absence of any interaction, $H_{0,\mathcal{R}}$ must not change the status of the observables of the reservoir.*

In the models discussed in this chapter, the only relevant *observables of the reservoir* are the number operators associated to it. For instance, in Section 3, we could introduce two such operators, $\int_{\mathbb{R}} A^\dagger(k)A(k)\,dk$ and $\int_{\mathbb{R}} B^\dagger(k)B(k)\,dk$, both commuting with the Hamiltonian H in (11.12) in absence of interactions, i.e., if we fix $\gamma_A = \gamma_B = \lambda = 0$.

The final ingredient in (11.7) is now $H_{\mathcal{S},\mathcal{R}}$. Its analytic expression is fixed by the simplest possible requirement, which is a simple extension of rule **R3**, and which is given by the following rule:

R5:– *in the interaction Hamiltonian $H_{\mathcal{S},\mathcal{R}}$ each annihilation operator of the system is linearly coupled to a corresponding creation operator of the reservoir as in (11.6), and viceversa.*

This is exactly what happens, for instance, in the damped love affair, described in (11.12) of Section 3, where Alice's and Bob's annihilation and creation operators are linearly coupled to the creation and annihilation operators of Alice's and Bob's reservoirs.

One might wonder why this particularly simple (i.e., linear) form of the interaction is assumed. The reason is very simple: because it works! What we need here is to produce damping, and this simple choice of $H_{\mathcal{S},\mathcal{R}}$ produces indeed damping, with not many additional analytical complications.

It is not hard to imagine that the five rules given in this chapter do not cover all possible situations. However, they already give some useful leading rules which can be adapted to several, quite different situations. Some of these applications will be reviewed in the rest of this chapter.

3 An *easy* dynamical system: love affairs

In [8] we have used the general framework introduced in Section 2 to describe a love relation between two (Alice and Bob)) or three (Alice, Bob and Carla) actors. In particular, the simplest hamiltonian we have adopted to describe the interaction between Alice and Bob is the following:

$$H = \lambda\left(a_1^M a_2^\dagger + \text{h.c.}\right), \tag{11.8}$$

where λ is the interaction parameter. Here a_j and $a_j^\dagger$ are two-modes annihilation and creation operators, $[a_j, a_k^\dagger] = \delta_{j,k}1\!\!1$. In agreement with what we have discussed in Section 2.1, the physical meaning of H can be deduced considering the action of, say, $a_1^M a_2^\dagger$ on the vector describing the system at time $t = 0$,

$\varphi_{n_1,n_2} = \frac{1}{\sqrt{n_1!n_2!}}(a_1^\dagger)^{n_1}(a_2^\dagger)^{n_2}\varphi_{0,0}$. This means that, at $t=0$, Bob is in the state n_1, i.e., n_1 is Bob's LoA, while Alice is in the state n_2. Now, because of the definition of φ_{n_1,n_2}, $a_1^M a_2^\dagger\varphi_{n_1,n_2}$, which is different from zero only if $M < n_1$, is proportional to φ_{n_1-M,n_2+1}. Hence, Bob's interest for Alice decreases of M units while Alice's interest for Bob increases of 1 unit. Of course, the Hamiltonian (11.8) also contains the opposite effect. Indeed, because of the presence of $a_2 a_1^{\dagger M}$ in H, if $n_2 \geq 1$ we see that $a_2 a_1^{\dagger M}\varphi_{n_1,n_2}$ is proportional to φ_{n_1+M,n_2-1}: hence, Bob's interest is increasing (of M units) while Alice looses interest in Bob. This is exactly in line with what Rule **R3** prescribes. It is clear that H has no free Hamiltonian H_0. This was, historically, the way in which the model was introduced originally, before the set of rules outlined above, and the importance of a free Hamiltonian, were fully understood. We decided to leave here the model as it was first defined also because H_0 will be considered in its *open system* version, see (3.5).

It is not hard to check that $I(t) := N_1(t) + MN_2(t)$ is a constant of motion: $I(t) = I(0) = N_1(0) + MN_2(0)$, for all $t \in \mathcal{R}$. This is a consequence of the following commutation result: $[H,I] = 0$. Therefore, during the time evolution, a certain *global attraction* between Alice and Bob is preserved and can only be exchanged between the two.

From the point of view of the differential equations of motion, they can be deduced using the Heisenberg rule $\dot{X}(t) = i[H,X(t)]$. In [8] it is shown that, if $M = 1$ (linear model), these equations can be solved analytically: calling $n_j(t) := \omega_{n_1,n_2}(N_j(t)) = \langle\varphi_{n_1,n_2}, N_j(t)\varphi_{n_1,n_2}\rangle$, $j = 1,2$, we find that

$$n_1(t) = n_1\cos^2(\lambda t) + n_2\sin^2(\lambda t), \qquad n_2(t) = n_2\cos^2(\lambda t) + n_1\sin^2(\lambda t), \tag{11.9}$$

so that, in particular, $\omega_{n_1,n_2}(I(t)) = n_1 + n_2$ which is, as expected, constant in time. When $M > 1$ the equations are no longer linear, but still we can easily deduce numerical solutions, showing again an oscillating behavior. For instance, in Figure 11.1 we plot the LoA's of Alice and Bob taking $M = 2$, assuming that at $t = 0$ the two lovers are both in the second level ($n_1 = n_2 = 2$).

It is important to stress that, even if in principle the Hilbert space is infinite-dimensional, the existence of an integral of motion makes it *effectively finite-dimensional*. This is because $N_1(t) + MN_2(t)$ must be constant in time, so that the eigenvectors φ_{n_1,n_2} involved in the description of $\mathcal{S}$ cannot have arbitrarily large n_1 and n_2: in fact, their maximum values are necessarily restricted by the value of $N_1(0) + MN_2(0)$. Hence, not all the *energy levels* of the system can be occupied.

In [8] we have also discussed a generalization of the model based on the existence of a third actor, Carla, who is Bob's lover. The situation can be summarized as follows:

1. Bob can interact with both Alice and Carla, but Alice (respectively, Carla) does not suspect of Carla's (respectively, Alice's) role in Bob's life;

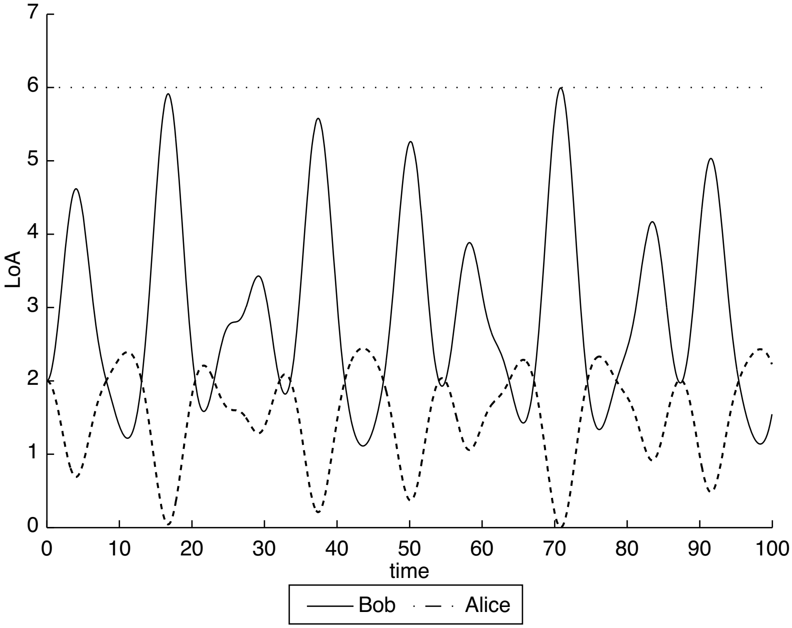


Figure 11.1 Alice's and Bob's LoA's vs. time with initial conditions $(2,2)$ and $M=2$

2. if Bob's LoA for Alice increases then Alice's LoA for Bob decreases and viceversa;
3. if Bob's LoA for Carla increases then Carla's LoA for Bob decreases and viceversa;
4. if Bob's LoA for Alice increases then his LoA for Carla decreases (not necessarily by the same amount) and viceversa.

The Hamiltonian of the system is assumed to be

$$H=\lambda_{12}\left((a_{12}^\dagger)^{M_{12}}\,a_2+a_{12}^{M_{12}}\,a_2^\dagger\right)+\lambda_{13}\left((a_{13}^\dagger)^{M_{13}}\,a_3+a_{13}^{M_{12}}\,a_3^\dagger\right)+\lambda_1\left(a_{12}^\dagger a_{13}+a_{12}a_{13}^\dagger\right), \tag{11.10}$$

where again $[a_\alpha,a_\beta^\dagger]=\delta_{\alpha,\beta}1\!\!1$, $\alpha,\beta=12,13,2,3$, the other commutators being zero. Again, $H_0=0$ here. This Hamiltonian is particularly easy to handle if $M_{12}=M_{13}=1$, since in this case the differential equations of motion become linear, but we will not make this assumption here. Also in this case an integral of motion does exist, and looks like

$$J:=N_{12}+N_{13}+M_{12}N_2+M_{13}N_3,$$

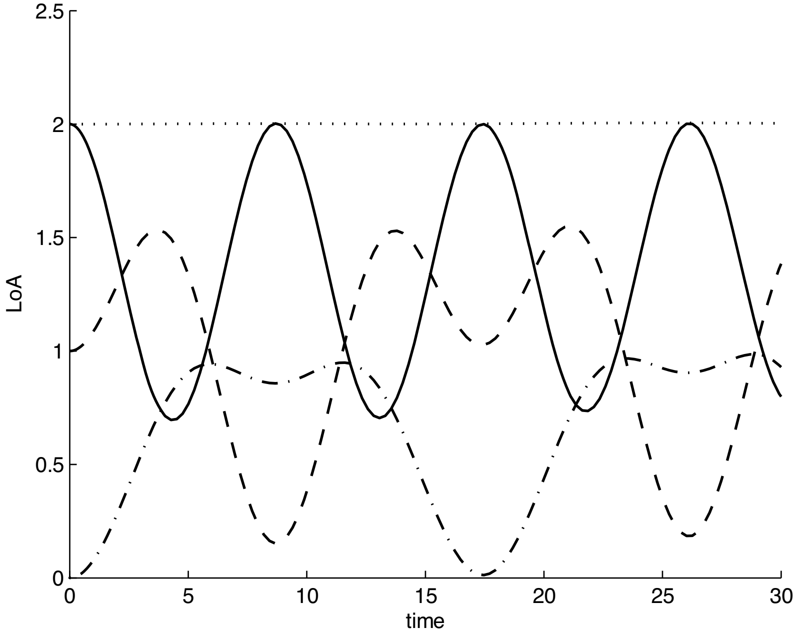


Figure 11.2 $M_{12} = 1$, $M_{13} = 2$: LoA vs. time of: Bob vs. Alice (continuous line), Bob vs. Carla (dashed line), Alice (dashed–dotted line), Carla (dotted line) with initial condition $(2,1,0,2)$. Periodic behaviors are observed

where, as usual, $N_\alpha = a_\alpha^\dagger a_\alpha$. The equations of motion are,

$$\begin{cases} i\dot{a}_{12}(t) = \lambda_{12} M_{12}\,(a_{12}^\dagger(t))^{M_{12}-1} a_2(t) + \lambda_1\, a_{13}(t), \\ i\dot{a}_{13}(t) = \lambda_{13} M_{13}\,(a_{13}^\dagger(t))^{M_{13}-1}\, a_3(t) + \lambda_1\, a_{12}(t), \\ i\dot{a}_2(t) = \lambda_{12}\, a_{12}(t), \\ i\dot{a}_3(t) = \lambda_{13}\, a_{13}(t), \end{cases} \tag{11.11}$$

which, as expected, are nonlinear and cannot be solved analytically unless if $M_{12} = M_{13} = 1$. However, numerical techniques can be used. For instance, if $M_{12} = 1$, $M_{13} = 2$, and if the initial conditions for N_{12}, N_{13}, N_2 and N_3 are $(2,1,0,2)$, we get Figure 11.2, which shows a rather regular behavior in spite of the (small) nonlinearity of the differential equations.

We see from our analytical and numerical results that oscillations appear to be inevitable in these models, and in fact, this is probably so. Then, the lovers cannot reach any equilibrium whatsoever. This is a rather unpleasant feature in any realistic love story. However, this can be avoided and, in fact, we have proposed a possible way out in [9], which will be important also for other applications, as we will discuss later on: an equilibrium can be reached if Alice and Bob are forced to live in a *large world,* i.e., using different words, if they can interact also with other people other than among themselves. In this case, clearly, the Hamiltonian must be more complicated. Following our previous Rules **R1-R5** we define

$$\begin{cases} H = H_A + H_B + \lambda H_I, \\ H_A = \omega_a a^\dagger a + \int_{\mathbb{R}} \Omega_A(k) A^\dagger(k) A(k)\, dk + \gamma_A \int_{\mathbb{R}} \left(a^\dagger A(k) + a A^\dagger(k)\right) dk, \\ H_B = \omega_b b^\dagger b + \int_{\mathbb{R}} \Omega_B(k) B^\dagger(k) B(k)\, dk + \gamma_B \int_{\mathbb{R}} \left(b^\dagger B(k) + b B^\dagger(k)\right) dk, \\ H_I = a^\dagger b + a b^\dagger. \end{cases} \tag{11.12}$$

All the constant in (11.12) are real quantities, and the following bosonic commutation rules are assumed:

$$[a, a^\dagger] = [b, b^\dagger] = 1\!\!1, \quad [A(k), A^\dagger(q)] = [B(k), B^\dagger(q)] = 1\!\!1 \delta(k - q), \tag{11.13}$$

while all the other commutators are zero. H_A and H_B respectively describe the interaction of Alice and Bob with their own reservoirs, which consist of several (infinite) ingredients. In this particular case, for two-actors quadratic Hamiltonian, the dynamical behavior can be deduced analytically and it looks like

$$n_a(t) = e^{-2\pi \gamma_A^2 t/\Omega_A} \left(n_a \cos^2(\lambda t) + n_b \sin^2(\lambda t)\right),$$

$$n_b(t) = e^{-2\pi \gamma_A^2 t/\Omega_A} \left(n_b \cos^2(\lambda t) + n_a \sin^2(\lambda t)\right),$$

which produce damped oscillations for both Alice and Bob. The speed of decay of their LoA is related to γ_A^2/Ω_A which, in our working conditions, see [9], coincides with γ_B^2/Ω_B. In particular, the stronger the interaction between, say, Alice and her reservoir, the faster the decay to zero of her love for Bob. In order to somehow stabilize the two lovers, we need this ratio to be very small, so that, even if the LoA's go both to zero, these process is very slow, allowing a sort of *decent love story* between Alice and Bob. On the other hand, if this ratio is very high, convergence to zero of $n_a(t)$ and $n_b(t)$ is very fast: Alice and Bob are going to split soon!

4 A *difficult* application: stock markets

The analysis carried out in the previous section is important to discuss, in a reasonably simple situation, how our general settings work like. The next step consists in applying the same ideas to an extremely more complicated problem, the description of some realistic, but still oversimplified stock market.

The basic assumptions used here to construct our first version of a SSM are the following:

1. the market consists of L traders exchanging a single kind of share;
2. the total number of shares, N, and the total amount of cash, K, are fixed in time;

3. a trader can only interact with a single other trader: i.e., each trader feels only a *two-body interaction*;
4. the traders can only buy or sell one share (or one block of shares) in any single transaction;
5. the price of the share (or of the block of shares) changes discontinuously, with discrete steps, multiples of a given monetary unit. We make no difference between the bid and the ask prices;
6. when the overall tendency of the market to sell a share, i.e., what we call the *market supply*, increases, then the price of the share decreases, and viceversa;
7. the market supply is expressed in term of natural numbers;
8. to simplify the notation, we fix the monetary unit to be equal to one.

Some of these assumptions will be relaxed later. However, for the moment, we will assume them to have a reasonably simple dynamical system. Of course Assumption 4. does not prevent at all the possibility of two traders to buy or sell more than one share (or block of shares). The only point is that the two traders must interact more than once. This is technically useful to avoid having a strongly non-linear model, with all the extra numerical and analytical difficulties that this would imply. Assumption 2. is a first consequence of the fact that the market is closed (the cash and the shares are neither created nor destroyed). The existence of these two conserved quantities will be used, together with Rules **R1**-**R3**, as a guideline to build up the Hamiltonian of the SSM. In fact, the two related operators, see $\hat{K}$ and $\hat{N}$ in formula (11.19) below must commute with this Hamiltonian. Assumption 3 means that it is more likely having two rather than three traders interacting simultaneously. Assumption 5. simply confirms the discrete nature of the model. Assumption 8 is used just to simplify the notation. Assumptions 6 and 7 provide a very simple mechanism to fix the value of the shares in terms of a global quantity, the market supply, which is a measure of the will of the various traders of the market to buy or sell the shares.[1] Of course, there is no problem in assuming that the supply is measured in terms of natural numbers, even because this is a very natural choice in our settings, much more than requiring that it changes continuously.

Looking at the above assumptions we immediately notice that some standard peculiarities of what in the literature is usually called a stock market, [3], are missing. For instance, we have not considered here any financial derivative, which, on the other hand, are the main interest in the monograph [3], which shares with this book the use of a typical quantum tool, the path integral, in the economical context. Also, as already mentioned, we will make no difference between the *bid* and *ask* prices. Ours is just a first step toward real systems, but, in our opinion, it gives already interesting and non trivial results.

The *formal* Hamiltonian of the model is the following operator:

$$\left\{\begin{array}{l} \tilde{H}=H_0+\tilde{H}_I, \text{ where} \\ H_0=\sum_{l=1}^L \alpha_l a_l^\dagger a_l+\sum_{l=1}^L \beta_l c_l^\dagger c_l+o^\dagger o+p^\dagger p \\ \tilde{H}_I=\sum_{i,j=1}^L p_{ij}\left[\left(a_i^\dagger c_i^{\hat{P}}\right)\left(a_j c_j^{\dagger^{\hat{P}}}\right)+\left(a_j^\dagger c_j^{\hat{P}}\right)\left(a_i c_i^{\dagger^{\hat{P}}}\right)\right]+(o^\dagger p+p^\dagger o), \end{array}\right. \tag{11.14}$$

where $\hat{P}=p^\dagger p$ and the following CCR are assumed:

$$[a_l,a_n^\dagger]=[c_l,c_n^\dagger]=\delta_{ln}1\!\!1, \qquad [p,p^\dagger]=[o,o^\dagger]=1\!\!1. \tag{11.15}$$

All the other commutators are zero. The quantities α_l, β_l, p_{ij} and so on are real numbers. In particular, p_{ij} can only be one or zero, depending on the fact that τ_i interacts or not with τ_j. Of course, $p_{ii}=0$. The operators $(a_l,a_l^\dagger)$ and $(c_l,c_l^\dagger)$ modify (by acting on a suitable vector) respectively the number of shares and the units of cash in the portfolio of the trader τ_l. The operators $(p,p^\dagger)$ change the price of the shares, while $(o,o^\dagger)$ change the value of the market supply. Of course, these changes are positive or negative, depending on whether creation or annihilation operators are used. The vector states of the market are defined in the usual way:

$$\omega_{\{n\};\{k\};O;M}(X)=<\varphi_{\{n\};\{k\};O;M},X\varphi_{\{n\};\{k\};O;M}>, \tag{11.16}$$

where $\{n\}=n_1,n_2,\dots,n_L$ and $\{k\}=k_1,k_2,\dots,k_L$ describe the number of shares and the units of cash of each trader at $t=0$, while O and M fix the initial values of the market supply and of the value of the shares. X is a generic observable of the SSM. More explicitly

$$\varphi_{\{n\};\{k\};O;M}:=\frac{(a_1^\dagger)^{n_1}\cdots(a_L^\dagger)^{n_L}(c_1^\dagger)^{k_1}\cdots(c_L^\dagger)^{k_L}(o^\dagger)^O(p^\dagger)^M}{\sqrt{n_1!\dots n_L!k_1!\dots k_L!O!M!}}\varphi_{\mathbf{0}}. \tag{11.17}$$

Here $\varphi_{\mathbf{0}}$ is the *vacuum* of the model: $a_j\varphi_{\mathbf{0}}=c_j\varphi_{\mathbf{0}}=p\varphi_{\mathbf{0}}=o\varphi_{\mathbf{0}}=0$, for $j=1,2,\dots,L$, and n_j, k_j, O and M are natural numbers.

The interpretation of the Hamiltonian is the key element in our approach, and follows from the general ideas discussed before: first of all, see Rule **R1**, H_0 is the free Hamiltonian of the system, which contains no interaction between the various ingredients of the market. It is clear that H_0 commutes with all the observables of the market, i.e., with all the *number operators* relevant for the description of our SSM. Concerning $\tilde{H}_I$, this has been written using an extended version of Rule **R3**: the term $o^\dagger p$ is responsible for the supply to go up and for a simultaneous lowering of the price of the shares. This is not very different from what happened, for instance, in Section 3, in our (linear) description of love affairs, but with a completely different interpretation. Moreover, because of $\left(a_i^\dagger c_i^{\hat{P}}\right)\left(a_j c_j^{\dagger^{\hat{P}}}\right)$, trader τ_i increases of one unit the number of shares in his portfolio but, at the same time, his cash decreases, because of $c_i^{\hat{P}}$, of as many

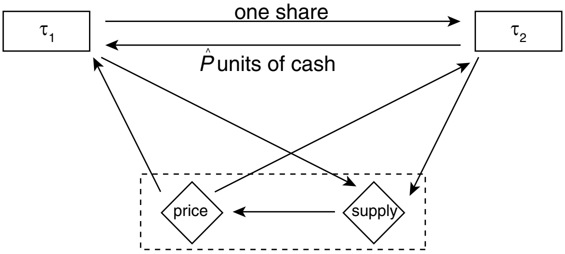


Figure 11.3 A schematic view of a two-trader market

units of cash as the price operator $\hat{P}$ demands. Clearly, trader τ_j behaves in the opposite way: he loses one share because of a_j but his cash increases because of $(c_j^\dagger)^{\hat{P}}$. Hence the meaning of $\tilde{H}$ in (11.14), and of $\tilde{H}_I$ in particular, is clear: it describes a SSM where two traders may buy or sell one share in each transaction, earning or paying money in this operation, and in which the price of the shares is related to the value of the supply operator as prescribed by Assumption 6. A schematic view of this market is given in Figure 11.3, where we consider (just) two traders, exchanging shares and money on the top while, at the bottom, the mechanism which fixes the price is shown, with the upwards arrows carrying the information of the value of the shares to the traders. The two downwards arrows represent the *feelings* of the various traders which, together, *construct* the market supply.

In the Hamiltonian in (11.14) a mathematical problem in its definition is somehow hidden: since c_j and $c_j^\dagger$ are not self-adjoint operators, it is not obvious at a first sight how to define the operators $c_j^{\hat{P}}$ and $(c_j^\dagger)^{\hat{P}}$: they look like non self-adjoint, unbounded, operators raised to some power, $\hat{P}$, which, by itself, is a different unbounded operator. However, from an *economical* point of view, $\tilde{H}$ is perfectly reasonable. One possible way out from this problem consists in replacing $\tilde{H}$ with an *effective* Hamiltonian, H, defined as

$$\left\{\begin{array}{l} H = H_0 + H_I, \text{ where} \\ H_0 = \sum_{l=1}^L \alpha_l a_l^\dagger a_l + \sum_{l=1}^L \beta_l c_l^\dagger c_l + o^\dagger o + p^\dagger p \\ H_I = \sum_{i,j=1}^L p_{ij} \left[\left(a_i^\dagger c_i^M\right)\left(a_j {c_j^\dagger}^M\right) + \left(a_j^\dagger c_j^M\right)\left(a_i {c_i^\dagger}^M\right)\right] + (o^\dagger p + p^\dagger o), \end{array}\right. \tag{11.18}$$

where M could be taken to be some *average value* of the price operator, $<\hat{P}>$. In this section, we will take M as the initial value of the share, and we will motivate this choice later on. This technique is quite diffused in quantum mechanics, where one replaces the *original Hamiltonian*, physically motivated but producing very difficult Heisenberg equations of motion, with an *effective Hamiltonian* which, in principle, still describes most of the original features of the system and whose related equations are easier to be solved and mathematically less *problematic*. Replacing $\hat{P}$ with M, however, means that we are essentially *freezing* the price of our share, removing one of the (essential) degrees of freedom of the system out of our market. Moreover, since $\hat{P}$ is also related to $\hat{O} = o^\dagger o$ by an integral of motion, see (11.19) below, this also means that we are *effectively* removing also a second degree of freedom from the market, keeping as the only relevant variables of the SSM the shares and the cash. However, this is only partially true, since the term $o^\dagger p + p^\dagger o$ in H_I is still there and produces, as we will see in a moment, a simple but not entirely trivial dynamics for $\hat{P}(t)$ and $\hat{O}(t)$.

Three integrals of motion for our model trivially exist:

$$\hat{N} = \sum_{i=1}^{L} a_i^\dagger a_i, \quad \hat{K} = \sum_{i=1}^{L} c_i^\dagger c_i \quad \text{and} \quad \hat{\Gamma} = o^\dagger o + p^\dagger p. \tag{11.19}$$

This can be easily checked since the CCR in (11.15) imply that

$$[H,\hat{N}] = [H,\hat{\Gamma}] = [H,\hat{K}] = 0.$$

The fact that $\hat{N}$ is conserved clearly means that no new shares are introduced in the market and that no share is removed from the market. Of course, also the total amount of money must be a constant of motion since the cash is assumed to be used only to buy shares. Moreover, since also $\hat{\Gamma}$ commutes with H, then, if the mean value of $o^\dagger o$ increases with time, the mean value of the price operator $\hat{P} = p^\dagger p$ must decrease and viceversa. Of course, since going from $\tilde{H}$ to H $\hat{P}(t)$ is replaced by M, $\hat{O}(t)$ will be constant as well. Moreover, also the following operators commute with H and, as a consequence, are independent of time:

$$\hat{Q}_j = a_j^\dagger a_j + \frac{1}{M} c_j^\dagger c_j, \tag{11.20}$$

for $j = 1,2,\ldots,L$. It is important to stress that the Q_j's are no longer integrals of motion for the original Hamiltonian, where $\hat{P}$ is not yet replaced by M.

4.1 The thermodynamical limit

Here we concentrate on what we call the *semiclassical thermodynamical limit* of the model, i.e., a non-trivial suitable limit which can be deduced for a very large number of traders, that is when $L \to \infty$.

In this case our model is defined by the Hamiltonian in (11.18), but with $M = 1$. From an *economical* point of view, this is not a major requirement, since it simply corresponds to fixing the price of the share to one: if you buy a share, then your liquidity decreases of one unit while it increases, again of one unit, if you are selling that share. Needless to say, this is strongly related to the fact that the original time-dependent price operator $\hat{P}(t)$ has been replaced by its mean value, M.

Of course, having $M = 1$ does not modify the integrals of motion found before: $\hat{N}$, $\hat{K}$ and $Q_j = \hat{n}_j + \hat{k}_j$, $j = 1,2,\ldots,L$, as well as $\hat{\Gamma} = \hat{O} + \hat{P}$. They all commute with H, which we now write as

$$\begin{cases} H = h + h_{po}, \text{ where} \\ h = \sum_{l=1}^{L} \alpha_l \hat{n}_l + \sum_{l=1}^{L} \beta_l \hat{k}_l + \sum_{i,j=1}^{L} p_{ij} \left[\left(a_i^\dagger c_i \right) \left(a_j c_j^\dagger \right) + \left(a_j^\dagger c_j \right) \left(a_i c_i^\dagger \right) \right] \\ h_{po} = o^\dagger o + p^\dagger p + (o^\dagger p + p^\dagger o). \end{cases} \tag{11.21}$$

Incidentally, also $\hat{\Delta} := \hat{O} - \hat{P}$ commutes with H. In our situation, the term h_{po} is unessential: it would cause changes in the price of the share and in the market supply, but these are frozen by our approximations. For this reason, from now on, we will identify H only with h in (11.21) and we will work only with this Hamiltonian. Let us introduce the operators

$$X_i = a_i c_i^\dagger, \tag{11.22}$$

$i = 1, 2, \ldots, L$. This is (a first version of) what we call *the selling operator*: it acts on a state of the market destroying a share and creating one unit of cash in the portfolio of the trader τ_i. Its adjoint $X_i^\dagger = a_i^\dagger c_i$, for obvious reasons, is called *the buying operator*. The Hamiltonian h can be rewritten as

$$h = \sum_{l=1}^{L} \left(\alpha_l \hat{n}_l + \beta_l \hat{k}_l \right) + \sum_{i,j=1}^{L} p_{ij} \left(X_i^\dagger X_j + X_j^\dagger X_i \right). \tag{11.23}$$

The following commutation relations can be deduced by the CCR in (11.15):

$$[X_i, X_j^\dagger] = \delta_{ij}(\hat{k}_i - \hat{n}_i), \qquad [X_i, \hat{n}_j] = \delta_{ij} X_i \qquad [X_i, \hat{k}_j] = -\delta_{ij} X_i, \tag{11.24}$$

which show how the operators $\{\{X_i, X_i^\dagger, \hat{n}_i, \hat{k}_i\}, i = 1, 2, \ldots, L\}$ are closed under commutation relations. This is quite important, since, introducing the operators $X_l^{(L)} = \sum_{i=1}^{L} p_{li} X_i$, $l = 1, 2, \ldots, L$, we get the following system of differential equations:

$$\begin{cases} \dot{X}_l = i(\beta_l - \alpha_l) X_l + 2i X_l^{(L)} (2\hat{n}_l - Q_l), \\ \dot{\hat{n}}_l = 2i \left(X_l X_l^{(L)^\dagger} - X_l^{(L)} X_l^\dagger \right). \end{cases} \tag{11.25}$$

This system, as l takes all the values $1, 2, \ldots, L$, is a closed system of differential equations for which an unique solution surely exists. Unfortunately, this is simply an existence result, not very useful in practice. More concretely, system (11.25) can be solved by introducing the so-called *mean-field approximation*, which essentially consists in replacing p_{ij} with $\frac{\tilde{p}}{L}$, for some $\tilde{p} \geq 0$. This is a standard approximation in quantum many body, widely used in solid state and in statistical mechanics. After this replacement we have that

$$X_l^{(L)} = \sum_{i=1}^{L} p_{li} X_i \longrightarrow \frac{\tilde{p}}{L} \sum_{i=1}^{L} X_i,$$

whose limit, for L diverging, exists only in suitable topologies, [10, 11], like, for instance, the strong one restricted to a set of relevant states[2]. Let τ be such a topology. We define

$$X^{\infty} = \tau - \lim_{L\to\infty} \frac{\tilde{p}}{L} \sum_{i=1}^{L} X_i, \tag{11.26}$$

where, as it is clear, the dependence on the index l is lost because of the replacement $p_{li} \to \frac{\tilde{p}}{L}$. The operator X^{∞} commutes (in some weak sense, see [10]) with all the observables of our stock market: $[X^{\infty}, A] = 0$ for all oservable A. In this limit, the system in (11.25) can be rewritten as

$$\begin{cases} \dot{X}_l = i(\beta_l - \alpha_l)X_l + 2iX^{\infty}(2\hat{n}_l - Q_l), \\ \dot{\hat{n}}_l = 2i\left(X_l X^{\infty\dagger} - X^{\infty} X_l^{\dagger}\right), \end{cases} \tag{11.27}$$

which can be analytically solved easily, at least under the additional useful hypothesis, concerning the parameters of the free Hamiltonian:

$$\beta_l - \alpha_l =: \Phi \neq \nu, \tag{11.28}$$

for all $l = 1, 2, \ldots, L$ (but also in other and more general situations). Here $\nu = \Phi + 4\eta - 2Q$, where

$$\eta := \tau - \lim_{L\to\infty} \frac{1}{L} \sum_{i=1}^{L} \hat{n}_i, \qquad Q := \tau - \lim_{L\to\infty} \frac{1}{L} \sum_{i=1}^{L} \hat{Q}_i.$$

We refer to [12] for the details of this derivation and for further generalizations. Here we just write the final result, which, calling as usual $n_l(t) = \omega_{\{n\};\{k\};O;M}(\hat{n}_l(t))$, is

$$n_l(t) = \frac{1}{\omega^2} \left\{ n_l(\Phi - \nu)^2 - 8|X_0^{\infty}|^2 \left(k_l(\cos(\omega t) - 1) - n_l(\cos(\omega t) + 1)\right) \right\}, \tag{11.29}$$

where $\omega = \sqrt{(\Phi - \nu)^2 + 16|X_0^{\infty}|^2}$. This, and the existence of the various integral of motion, allows also to find the time evolution for the portfolio of each trader, which we define as $\pi_l(t) = k_l(t) + Mn_l(t)$. Hence, at least in our assumptions, these portfolios are under control. Of course, the periodic behavior of $n_l(t)$ is transferred to $\pi_l(t)$, and this gives a measure of the limits of the model presented here, since it is hard to believe that the portfolios of the traders in a real market may simply change periodically! However, in several recent applications of quantum mechanics to markets, periodic behaviors of some kind are often deduced. For instance, in [13], a periodic behavior is deduced in connection with the Chinese market. Also, in [14], Schaden suggests that a periodic behavior is, in a sense, unavoidable anytime we deal with a closed market with few traders, while a non periodic behavior (some decay, for instance) can be obtained only in presence of many traders.

4.2 A time dependent point of view

In this section we will extend our model by introducing first several kind of shares and we will adopt a slightly different point of view. In particular, our Hamiltonian will have no H_{price} contribution at all, since the price operators $\hat{P}_\alpha$, $\alpha = 1,\dots,L$, will here be replaced from the very beginning by external classical fields $P_\alpha(t)$, whose time dependence describes, as an input of the model fixed by empirical data, the variation of the prices of the shares. This implies that any change of the prices is automatically included in the model through the analytic expressions of the functions $P_\alpha(t)$. Hence the interaction Hamiltonian H_I turns out to be a time-dependent operator, $H_I(t)$. In more detail, the Hamiltonian is $H(t) = H_0 + \lambda H_I(t)$, which we can write, introducing the time-depending *selling* and *buying* operators which extend those introduced before,

$$x_{j,\alpha}(t) := a_{j,\alpha}\, c_j^{\dagger P_\alpha(t)}, \qquad x^\dagger_{j,\alpha}(t) := a^\dagger_{j,\alpha}\, c_j^{P_\alpha(t)}, \tag{11.30}$$

as

$$H(t) = \sum_{j,\alpha} \omega_{j,\alpha}\, \hat{n}_{j,\alpha} + \sum_j \omega_j \hat{k}_j + 2\lambda \sum_{i,j,\alpha} p^{(\alpha)}_{i,j} x^\dagger_{i,\alpha}(t) x_{j,\alpha}(t), \tag{11.31}$$

where $H_I(t) = 2\sum_{i,j,\alpha} p^{(\alpha)}_{i,j} x^\dagger_{i,\alpha}(t) x_{j,\alpha}(t)$ and $H_0 = \sum_{j,\alpha} \omega_{j,\alpha}\, \hat{n}_{j,\alpha} + \sum_j \omega_j \hat{k}_j$. This is not very different from the Hamiltonian in (11.23), as one can see.

Concerning the notation, while the latin indexes are related to the traders, the greek ones appear here because we are now considering different type of shares in our SSM. The related L price functions $P_\alpha(t)$ will be taken piecewise constant, since it is quite natural to assume that the price of a share changes discontinuously: it has a certain value before the transaction and (in general) a different value after the transaction. Furthermore, this new value does not change until the next transaction takes place. To be specific, we introduce a time step h which we might call *the time of transaction*, and we divide the interval $[0,t[$ in subintervals which, for simplicity, we consider having the same duration h: $[0,t[\,=\, [t_0,t_1[\,\cup\, [t_0,t_1[\,\cup\, [t_1,t_2[\cdots [t_{M-1},t_M[$, where $t_0 = 0$, $t_1 = h$, $\dots$, $t_{M-1} = (M-1)h = t-h$, $t_M = Mh = t$. Hence $h = t/M$. As for the prices, we put

$$P_\alpha(t) = \begin{cases} P_{\alpha,0}, & t \in [t_0,t_1[, \\ P_{\alpha,1}, & t \in [t_1,t_2[, \\ \dots\dots, & \\ P_{\alpha,M-1}, & t \in [t_{M-1},t_M[, \end{cases} \tag{11.32}$$

for $\alpha = 1,\dots L$. An o.n. basis in the Hilbert space $\mathcal{H}$ of the model is now the set of vectors defined as

$$\varphi_{\{n_{j,\alpha}\};\{k_j\}} := \frac{a_{1,1}^{\dagger\ n_{1,1}} \cdots a_{N,L}^{\dagger\ n_{N,L}}\, c_1^{\dagger k_1} \cdots c_N^{\dagger k_N}}{\sqrt{n_{11}! \cdots n_{N,L}! k_1! \cdots k_L!}}\, \varphi_{\mathbf{0}}, \tag{11.33}$$

where $\varphi_{\mathbf{0}}$ is the vacuum of all the annihilation operators involved here. To simplify the notation we introduce a set $\mathcal{F} = \{\{n_{j,\alpha}\};\{k_j\}\}$, so that the vectors of the basis will be simply written as $\varphi_{\mathcal{F}}$. The difference between these and the vectors in (11.17) stands clearly in the absence of the *quantum numbers O* and M. This is clearly due to the fact that $P(t)$ is no longer a degree of freedom of the system, and that the market supply does not even exist anymore, here.

Suppose now that at $t=0$ the market is described by a vector $\varphi_{\mathcal{F}_0}$. This means that, since $\mathcal{F}_0=\{\{n_{j,\alpha}^o\},\{k_j^o\}\}$, at $t=0$ the trader τ_1 has $n_{1,1}^o$ shares of the first type, $n_{1,2}^o$ shares of the second type, …, and k_1^o units of cash. Analogously, the trader τ_2 has $n_{2,1}^o$ shares of the first type, $n_{2,2}^o$ shares of the second type, …, and k_2^o units of cash. And so on. We want to compute the probability that at time t the market has moved to the configuration $\mathcal{F}_f=\{\{n_{j,\alpha}^f\},\{k_j^f\}\}$. This means that, for example, τ_1 has now n_{11}^f shares of the first type, n_{12}^f shares of the second type, …, and k_1^f units of cash.

Similar problems arise quite often in ordinary quantum mechanics: we need to compute a probability transition from the original state $\varphi_{\mathcal{F}_0}$ to a final state $\varphi_{\mathcal{F}_f}$, and this is the reason why we will use here a somehow standard time-dependent perturbation scheme for which we refer to [15]. The main difference with respect to what we have done so far in this chapter is that we adopt now the Schrödinger rather than the Heisenberg picture since it is more convenient in this kind of computations. Hence the market is described by a time-dependent wave function $\Psi(t)$ which, for $t=0$, reduces to $\varphi_{\mathcal{F}_0}$: $\Psi(0)=\varphi_{\mathcal{F}_0}$. The transition probability we are looking for is

$$P_{\mathcal{F}_0\rightarrow\mathcal{F}_f}(t):=\left|<\varphi_{\mathcal{F}_f},\Psi(t)>\right|^2. \tag{11.34}$$

The computation of $P_{\mathcal{F}_0\rightarrow\mathcal{F}_f}(t)$ goes like this: since the set of the vectors $\varphi_{\mathcal{F}}$ is an o.n basis in $\mathcal{H}$, the wave function $\Psi(t)$ can be written as

$$\Psi(t)=\sum_{\mathcal{F}}c_{\mathcal{F}}(t)e^{-iE_{\mathcal{F}}t}\varphi_{\mathcal{F}}, \tag{11.35}$$

where $E_{\mathcal{F}}$ is the eigenvalue of H_0 defined as

$$H_0\varphi_{\mathcal{F}}=E_{\mathcal{F}}\varphi_{\mathcal{F}}, \qquad \text{where} \qquad E_{\mathcal{F}}=\sum_{j,\alpha}\omega_{j,\alpha}n_{j,\alpha}+\sum_j\omega_jk_j. \tag{11.36}$$

This is a consequence of the fact that $\varphi_{\mathcal{F}}$ in (11.33) is an eigenstate of H_0, with eigenvalue. $E_{\mathcal{F}}$, which we call the free energy of $\varphi_{\mathcal{F}}$. Putting (11.35) in (11.34), and recalling that, if the eigenstates are not-degenerate,[3] the corresponding eigenvectors are orthogonal, $<\varphi_{\mathcal{F}},\varphi_{\mathcal{G}}>=\delta_{\mathcal{F},\mathcal{G}}$, we have

$$P_{\mathcal{F}_0\rightarrow\mathcal{F}_f}(t):=\left|c_{\mathcal{F}_f}(t)\right|^2 \tag{11.37}$$

The answer to our original question is therefore given if we are able to compute $c_{\mathcal{F}_f}(t)$ in the expansion (11.35). Due to the analytic form of our Hamiltonian, this cannot be done exactly. We adopt here a simple perturbation expansion in the interaction parameter λ appearing in the Hamiltonian H in (11.31), which we are assuming sufficiently small. In other words, we look for the coefficients in (11.35) having the form

$$c_{\mathcal{F}}(t) = c_{\mathcal{F}}^{(0)}(t) + \lambda c_{\mathcal{F}}^{(1)}(t) + \lambda^2 c_{\mathcal{F}}^{(2)}(t) + \cdots \tag{11.38}$$

Each $c_{\mathcal{F}}^{(j)}(t)$ of this expansion satisfies a differential equation, see [6, 15]:

$$\begin{cases} \dot{c}_{\mathcal{F}'}^{(0)}(t) = 0, \\ \dot{c}_{\mathcal{F}'}^{(1)}(t) = -i \sum_{\mathcal{F}} c_{\mathcal{F}}^{(0)}(t) e^{i(E_{\mathcal{F}'} - E_{\mathcal{F}})t} < \varphi_{\mathcal{F}'}, H_I(t) \varphi_{\mathcal{F}} >, \\ \dot{c}_{\mathcal{F}'}^{(2)}(t) = -i \sum_{\mathcal{F}} c_{\mathcal{F}}^{(1)}(t) e^{i(E_{\mathcal{F}'} - E_{\mathcal{F}})t} < \varphi_{\mathcal{F}'}, H_I(t) \varphi_{\mathcal{F}} >, \\ \ldots\ldots\ldots, \end{cases} \tag{11.39}$$

The first equation, together with the initial condition $\Psi(0) = \varphi_{\mathcal{F}_0}$, gives $c_{\mathcal{F}'}^{(0)}(t) = c_{\mathcal{F}'}^{(0)}(0) = \delta_{\mathcal{F}', \mathcal{F}_0}$. When we replace this solution in the differential equation for $c_{\mathcal{F}'}^{(1)}(t)$ we get, recalling again that $\Psi(0) = \varphi_{\mathcal{F}_0}$,

$$c_{\mathcal{F}'}^{(1)}(t) = -i \int_0^t e^{i(E_{\mathcal{F}'} - E_{\mathcal{F}_0})t_1} < \varphi_{\mathcal{F}'}, H_I(t_1) \varphi_{\mathcal{F}_0} > dt_1, \tag{11.40}$$

at least if $\mathcal{F}_0 \neq \mathcal{F}'$. Using this in (11.39) we further get

$$c_{\mathcal{F}'}^{(2)}(t) = (-i)^2 \sum_{\mathcal{F}} \int_0^t \left(\int_0^{t_2} e^{i(E_{\mathcal{F}} - E_{\mathcal{F}_0})t_1} h_{\mathcal{F}, \mathcal{F}_0}(t_1) dt_1 \right) e^{i(E_{\mathcal{F}'} - E_{\mathcal{F}})t_2} h_{\mathcal{F}', \mathcal{F}}(t_2) dt_2, \tag{11.41}$$

where we have introduced the shorthand notation

$$h_{\mathcal{F}, \mathcal{G}}(t) := < \varphi_{\mathcal{F}}, H_I(t) \varphi_{\mathcal{G}} >. \tag{11.42}$$

Of course, higher order corrections could also be deduced simply by iterating this procedure. This might be relevant, for instance for not so small values of λ, but we will not consider it here.

4.2.1 First order corrections

We apply now this general procedure to the analysis of our stock market, by computing $P_{\mathcal{F}_0 \to \mathcal{F}_f}(t)$ in (11.37) up to the first order corrections in powers of λ and assuming that $\mathcal{F}_f$ is different from $\mathcal{F}_0$. Hence we have

$$P_{\mathcal{F}_0 \to \mathcal{F}_f}(t) = \left| c_{\mathcal{F}_f}^{(1)}(t) \right|^2 = \lambda^2 \left| \int_0^t e^{i(E_{\mathcal{F}_f} - E_{\mathcal{F}_0})t_1} h_{\mathcal{F}_f, \mathcal{F}_0}(t_1) dt_1 \right|^2. \tag{11.43}$$

Using (11.32) and introducing $\delta E = E_{\mathcal{F}_f} - E_{\mathcal{F}_0}$, after some simple computations we get

$$P_{\mathcal{F}_0 \to \mathcal{F}_f}(t) = \lambda^2 \left(\frac{\delta E\, h/2}{\delta E/2} \right)^2 \left| \sum_{k=0}^{M-1} h_{\mathcal{F}_f, \mathcal{F}_0}(t_k) e^{i t_k \delta E} \right|^2. \tag{11.44}$$

The matrix elements $h_{\mathcal{F}_f,\mathcal{F}_0}(t_k)$ can be easily computed. Indeed, because of some standard properties of the bosonic operators, we find that

$$a_{i,\alpha}^\dagger a_{j,\alpha}\, c_i^{P_{\alpha,k}}\, c_j^\dagger P_{\alpha,k}\,\varphi_{\mathcal{F}_0} = \Gamma_{i,j;\alpha}^{(k)}\varphi_{\mathcal{F}_{0,k}^{(i,j,\alpha)}},$$

where

$$\Gamma_{i,j;\alpha}^{(k)} := \sqrt{\frac{(k_j^o+P_{\alpha,k})!}{k_j^o!}\,\frac{k_i^o!}{(k_i^o-P_{\alpha,k})!}\,n_{j,\alpha}^o\,(1+n_{i,\alpha}^o)}, \tag{11.45}$$

and $\mathcal{F}_{0,k}^{(i,j,\alpha)}$ differs from $\mathcal{F}_0$ only for the following replacements: $n_{j,\alpha}^o \to n_{j,\alpha}^o - 1$, $n_{i,\alpha}^o \to n_{i,\alpha}^o + 1$, $k_j^o \to k_j^o + P_{\alpha,k}$, $k_i^o \to k_i^o - P_{\alpha,k}$. Notice that, in our computations, we are implicitly assuming that $k_i^o \geq P_{\alpha,k}$, for all i, k and α. This has to be so, since otherwise the trader τ_i would have not enough money to buy a share Σ_α.

We find that

$$h_{\mathcal{F}_f,\mathcal{F}_0}(t_k) = 2\sum_{i,j,\alpha} p_{i,j}^{(\alpha)}\,\Gamma_{i,j;\alpha}^{(k)} < \varphi_{\mathcal{F}_f}, \varphi_{\mathcal{F}_{0,k}^{(i,j,\alpha)}} >. \tag{11.46}$$

Of course, due to the orthogonality of the vectors $\varphi_{\mathcal{F}}$'s, the scalar product $< \varphi_{\mathcal{F}_f}, \varphi_{\mathcal{F}_{0,k}^{(i,j,\alpha)}} >$ is different from zero (and equal to one) if and only if $n_{j,\alpha}^f = n_{j,\alpha}^o - 1$, $n_{i,\alpha}^f = n_{i,\alpha}^o + 1$, $k_i^f = k_i^o - P_{\alpha,k}$ and $k_j^f = k_j^o + P_{\alpha,k}$, and, moreover, if all the other *new* and *old* quantum numbers coincide.

For concreteness sake we now consider two simple situations: in the first example below we just assume that the prices of the various shares do not change with t. In the second example we consider the case in which only few (i.e., 3) changes occur.

Example 1:– constant prices

Let us assume that, for all k and for all α, $P_{\alpha,k} = P_\alpha(t_k) = P_\alpha$. This means that $\Gamma_{i,j;\alpha}^{(k)}$, $\mathcal{F}_{0,k}^{(i,j,\alpha)}$ and the related vectors $\varphi_{\mathcal{F}_{0,k}^{(i,j,\alpha)}}$ do not depend on k. Therefore, $h_{\mathcal{F}_f,\mathcal{F}_0}(t_k)$ is also independent of k. After few computations we get

$$P_{\mathcal{F}_0\to\mathcal{F}_f}(t) = \lambda^2\left(\frac{\sin(\delta E t/2)}{\delta E/2}\right)^2 \left|h_{\mathcal{F}_f,\mathcal{F}_0}(0)\right|^2, \tag{11.47}$$

from which we deduce the following transition probability per unit of time:

$$p_{\mathcal{F}_0\to\mathcal{F}_f} = \lim_{t,\infty}\frac{1}{t}P_{\mathcal{F}_0\to\mathcal{F}_f}(t) = 2\pi\,\lambda^2\,\delta(E_{\mathcal{F}_f} - E_{\mathcal{F}_0})\left|h_{\mathcal{F}_f,\mathcal{F}_0}(0)\right|^2. \tag{11.48}$$

This formula shows that, in this limit, and in our approximations, a transition between the states $\varphi_{\mathcal{F}_0}$ and $\varphi_{\mathcal{F}_f}$ is possible only if these two states have the same free energy. Moreover, the presence of $h_{\mathcal{F}_f,\mathcal{F}_0}(0)$ in Formulas (11.47) and (11.48) shows that, at the order we are working here, a transition is possible only if $\varphi_{\mathcal{F}_0}$ does not differ from $\varphi_{\mathcal{F}_f}$ for more than one share in two of the $n_{j,\alpha}$'s and for more than P_α units of cash in two of the k_j's[4]. All the other transitions, for instance those in which the numbers of shares differ for more than one unit, are forbidden at this order in perturbation theory.

Example 2:– few changes in the price

Let us now fix $M = 3$. Formula (11.44) can be rewritten as

$$P_{\mathcal{F}_0\to\mathcal{F}_f}(t) = 4\lambda^2\left(\frac{\sin(\delta Eh/2)}{\delta E/2}\right)^2 |\sum_{i,j,\alpha} p_{i,j}^{(\alpha)}(\Gamma_{i,j;\alpha}^{(0)} < \varphi_{\mathcal{F}_f}, \varphi_{\mathcal{F}_{0,0}^{(i,j,\alpha)}} >$$

$$+\Gamma_{i,j;\alpha}^{(1)} < \varphi_{\mathcal{F}_f}, \varphi_{\mathcal{F}_{0,1}^{(i,j,\alpha)}} > e^{ih\delta E} + \Gamma_{i,j;\alpha}^{(2)} < \varphi_{\mathcal{F}_f}, \varphi_{\mathcal{F}_{0,2}^{(i,j,\alpha)}} > e^{2ih\delta E})|^2 \qquad (11.49)$$

The meaning of this formula is not very different from the one deduced in the previous example: it is not surprising that, in order to get something different, we need to go to higher orders in powers of λ.

Let us now see what can be said about the portfolio of the trader τ_l. Assuming that we know the initial state of the system, then we clearly know, in particular, the value of $\tau_l's$ portfolio at time zero (actually, we know the values of the portfolios of all the traders!): $\hat{\pi}_l(0) = \sum_{\alpha=1}^{L} P_\alpha(0)\hat{n}_{l,\alpha}(0) + \hat{k}_l(0)$. Formula (11.44) gives the transition probability from $\varphi_{\mathcal{F}_0}$ to $\varphi_{\mathcal{F}_f}$. This probability is just a single contribution in the computation of the transition probability from a given $\hat{\pi}_l(0)$ to a certain $\hat{\pi}_l(t)$, since the same value of the portfolio of the $l-th$ trader could be recovered in general, at time t, for very many different states $\varphi_{\mathcal{F}_f}$: all the sets $\mathcal{G}$ with the same values of $n_{l,\alpha}^f$ and k_l^f, and with any other possible choice of $n_{l',\alpha}^f$ and $k_{l'}^f$, $l' \neq l$, give rise to the same value of the portfolio for τ_l. Hence, if we call $\tilde{\mathcal{F}}$ the set of all these sets, we just have to sum up over all these different contributions:

$$P_{\hat{\pi}_l^o\to\hat{\pi}_l^f}(t) = \sum_{\mathcal{G}\in\tilde{\mathcal{F}}} P_{\mathcal{F}_0\to\mathcal{G}}(t). \qquad (11.50)$$

In this way the transition probability could be, at least formally, computed at the desired order in powers of λ.

We refer to [6] for some remark concerning the validity of the approximations discussed here.

4.2.2 Second order corrections

We now want to show what happens going to the next order in the perturbation expansion. For that, we begin by considering the easiest situation, i.e., the case of a time independent perturbation H_I: the prices are constant in time. Hence the integrals in Formula (11.41) can be easily computed and the result is the following:

$$c_{\mathcal{F}_f}^{(2)}(t) = \sum_{\mathcal{F}} h_{\mathcal{F}_f,\mathcal{F}}(0)h_{\mathcal{F},\mathcal{F}_0}(0)\mathcal{E}_{\mathcal{F},\mathcal{F}_0,\mathcal{F}_f}(t), \qquad (11.51)$$

where

$$\mathcal{E}_{\mathcal{F},\mathcal{F}_0,\mathcal{F}_f}(t) = \frac{1}{E_{\mathcal{F}} - E_{\mathcal{F}_0}} \left(\frac{e^{i(E_{\mathcal{F}_f} - E_{\mathcal{F}_0})t} - 1}{E_{\mathcal{F}_f} - E_{\mathcal{F}_0}} - \frac{e^{i(E_{\mathcal{F}_f} - E_{\mathcal{F}})t} - 1}{E_{\mathcal{F}_f} - E_{\mathcal{F}}} \right).$$

Recalling definition (11.42), we rewrite equation (11.51) as

$$c^{(2)}_{\mathcal{F}_f}(t) = \sum_{\mathcal{F}} < \varphi_{\mathcal{F}_f}, H_I \varphi_{\mathcal{F}} > < \varphi_{\mathcal{F}}, H_I \varphi_{\mathcal{F}_0} > \mathcal{E}_{\mathcal{F},\mathcal{F}_0,\mathcal{F}_f}(t),$$

which explicitly shows that, up to the second order in λ, transitions between states which differ, for example, for 2 shares are allowed: it is enough that some intermediate state $\varphi_{\mathcal{F}}$ differs for (e.g., plus) one share from $\varphi_{\mathcal{F}_0}$ and for (e.g., minus) one share from $\varphi_{\mathcal{F}_f}$.

If the $P_\alpha(t)$'s depend on time the situation is a bit more complicated but not essentially different. We refer the interested reader to [6], where these and many other aspects are discussed.

4.2.3 Feynman graphs

Following [15] we now try to connect the analytic expression of a given approximation of $c_{\mathcal{F}_f}(t)$ with some kind of *Feynman graph* in such a way that the higher orders could be easily deduced considering a certain set of rules which we will obviously call *Feynman rules*.

The starting point is given by the expressions (11.40) and (11.41) for $c^{(1)}_{\mathcal{F}_f}(t)$ and $c^{(2)}_{\mathcal{F}_f}(t)$, which is convenient to rewrite in the following form:

$$c^{(1)}_{\mathcal{F}_f}(t) = -i \int_0^t e^{iE_{\mathcal{F}_f} t_1} < \varphi_{\mathcal{F}_f}, H_I(t_1) \varphi_{\mathcal{F}_0} > e^{-iE_{\mathcal{F}_0} t_1}\, dt_1, \qquad (11.52)$$

and

$$c^{(2)}_{\mathcal{F}_f}(t) = (-i)^2 \sum_{\mathcal{F}} \int_0^t dt_2 \int_0^{t_2} dt_1\, e^{iE_{\mathcal{F}_f} t_2} < \varphi_{\mathcal{F}_f}, H_I(t_2) \varphi_{\mathcal{F}} > e^{-iE_{\mathcal{F}} t_2} \times$$
$$\times e^{iE_{\mathcal{F}} t_1} < \varphi_{\mathcal{F}}, H_I(t_1) \varphi_{\mathcal{F}_0} > e^{-iE_{\mathcal{F}_0} t_1}. \qquad (11.53)$$

The reason why this is so useful is that, as we will now sketch, the different ingredients needed to find the Feynman rules are now explicitly separated and, therefore, easily identified. A graphical way to describe $c^{(1)}_{\mathcal{F}_f}(t)$ is given in the figure below: at $t = t_0$ the state of the system is $\varphi_{\mathcal{F}_0}$, which evolves freely (and therefore $e^{-iE_{\mathcal{F}_0} t_1} \varphi_{\mathcal{F}_0}$ appears) until the interaction occurs, at $t = t_1$. After the interaction the system is moved to the state $\varphi_{\mathcal{F}_f}$, which evolves again freely (and therefore $e^{-iE_{\mathcal{F}_f} t_1} \varphi_{\mathcal{F}_f}$ appears, and the different sign in (11.52) is due to the anti-linearity of the scalar product in the first variable). The free evolutions are represented by the upward inclined arrows, while the interaction between the initial and the final states, $< \varphi_{\mathcal{F}_f}, H_I(t_1) \varphi_{\mathcal{F}_0} >$, is described by the horizontal wavy line in Figure 11.4. Obviously, since the interaction may occur at any time

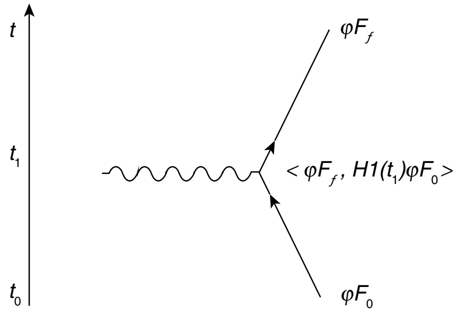


Figure 11.4 Graphical expression for $c^{(1)}_{\mathcal{F}_f}(t)$

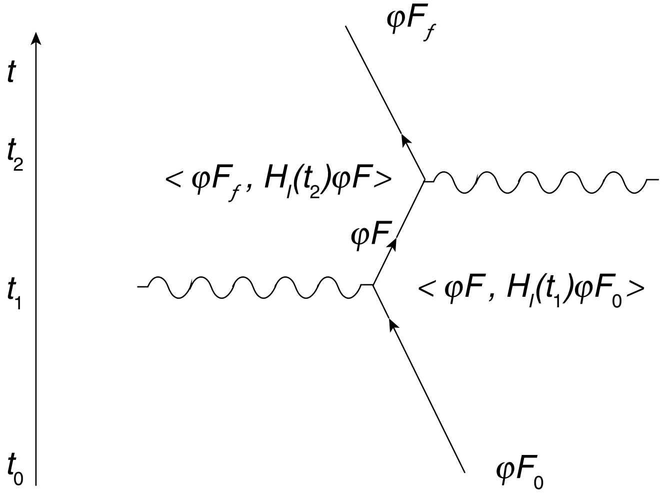


Figure 11.5 graphical expression for $c^{(2)}_{\mathcal{F}_f}(t)$

between 0 and t, we have to integrate on all these possible t_1's and multiply the result for $-i$, which is a sort of normalization constant.

In a similar way we can construct the Feynman graph for $c^{(2)}_{\mathcal{F}_f}(t)$, $c^{(3)}_{\mathcal{F}_f}(t)$ and so on. For example, $c^{(2)}_{\mathcal{F}_f}(t)$ can be deduced by a graph like the one in Figure 11.5, where two interactions occur, the first at $t = t_1$ and the second at $t = t_2$:

Because of the double interaction, we have now to integrate twice the result, recalling that $t_1 \in (0, t_2)$ and $t_2 \in (0, t)$. For the same reason we have to sum over all the possible intermediate states, $\varphi_{\mathcal{F}}$. The free time evolution for the various free fields also appear in Formula (11.53), as well as the normalization factor $(-i)^2$. Following these same rules we could also give a formal expression for the other coefficients, $c^{(3)}_{\mathcal{F}_f}(t)$, $c^{(4)}_{\mathcal{F}_f}(t)$ and so on: the third order correction $c^{(3)}_{\mathcal{F}_f}(t)$ contains, for instance, a double sum on the intermediate states, allowing in this way a transition from a state with, say, $n^o_{i,\alpha}$ shares to a state with $n^f_{i,\alpha}=n^o_{i,\alpha}+3$ shares, a triple time integral and a factor $(-i)^3$.

Summarizing we conclude that, at each order in perturbation theory, the contribution to the transition probability in (11.37) can be computed by considering first the relevant Feynman graph, computing the associated integral, and then using expansion (11.38).

5 The role of information

So far, in the analysis of our SSM, we have not considered any reservoir interacting with the traders. In fact, ours was a closed market, where cash and number of shares were preserved. We will now briefly discuss the effect of the outer world, focusing on what happens during the *preparation of the system*, i.e., while fixing the initial status of the various traders after they have been reached by some external information, but before they start to trade. For this reason, we consider two different time intervals: in the first one, $[0,t_1]$, the two traders, which are indistinguishable at $t=0$, receive a different amount of information. This allows them to react in different ways, so that, at time t_1, they are expected to be *different*. In this interval, our complete Hamiltonian, H_{full}, consists in a single *preparing term*, H. For $t>t_1$ to H is added a new interaction term: $H_{full}=H+\Theta(t-t_1)H_{ex}$, where $\Theta(t)=1$ if $t>0$, while $\Theta(t)=0$ otherwise. Since, in this paper, we will only be interested in the first time interval, $[0,t_1]$, the role of H_{ex} will not be very relevant in our analysis. However, we should mention that H_{ex} has been also considered, see [16], while what we are going to review here was originally discussed, together with other simpler models, in [17]. To simplify the treatment the price of the shares (just a single kind of shares!) will be fixed to be one.

The system we want to describe is made by just two traders,[5] τ_1 and τ_2, interacting with a source of information, $\mathcal{S}_{inf}$, described by the bosonic operators i_j, $i_j^\dagger$ and $\hat{I}_j=i_j^\dagger i_j$. $\mathcal{S}_{inf}$ interacts with the external world, mimicked by a reservoir described by the operators $r_j(k)$, $r_j^\dagger(k)$ and $\hat{R}_j(k)=r_j^\dagger(k)r_j(k)$, $j=1,2$, $k\in\mathbb{R}$. These operators are again, see below, bosonic. The reservoir is used to model the set of all the rumors, news, and external facts which, all together, create the final information. In fact, see below, the term $\gamma_j\int_{\mathbb{R}}(i_j^\dagger r_j(k)+i_jr_j^\dagger(k))dk$ is exactly what relates these reservoirs to what we can call *bad quality information* or, as we did in [17], *lack of information* (LoI).

The Hamiltonian is:

$$\begin{cases} H = H_0 + H_{int}, \\ H_0 = \sum_{j=1}^{2} (\omega_j^s \hat{S}_j + \omega_j^c \hat{K}_j + \Omega_j \hat{I}_j + \int_{\mathbb{R}} \Omega_j^{(r)}(k) \hat{R}_j(k)\, dk), \\ H_{int} = \sum_{j=1}^{2} \left[\lambda_{inf} \left(i_j(s_j^\dagger + c_j^\dagger) + i_j^\dagger (s_j + c_j) \right) + \gamma_j \int_{\mathbb{R}} (i_j^\dagger r_j(k) + i_j r_j^\dagger(k))\, dk \right], \end{cases} \tag{11.54}$$

where $\hat{S}_j = s_j^\dagger s_j$ and $\hat{K}_j = c_j^\dagger c_j$, and where the following CCRs are assumed,

$$[s_j, s_k^\dagger] = [c_j, c_k^\dagger] = [i_j, i_k^\dagger] = 1\!\!1 \delta_{j,k}, \quad [r_j(k), r_l^\dagger(q)] = 1\!\!1 \delta_{j,l} \delta(k-q),$$

all the other commutators being zero. This Hamiltonian is constructed following the Rules **R1-R5** discussed in Section 2.1. H contains a free *canonical* part H_0, which satisfies Rules **R1** and **R4**, while the two contributions in H_{int}, constructed according to Rules **R3** and **R5**, respectively describe: (i) the fact that when the LoI increases, the value of the portfolio decreases and vice versa; (ii) the fact that the LoI increases when the "value" of the reservoir decreases, and viceversa: for instance, the contribution $i_j r_j^\dagger(k)$ in H_{int} shows that the LoI decreases (so that the trader is *better informed*) when a larger amount of news, rumors, etc. reaches the trader. Notice also that, as anticipated, no interaction between τ_1 and τ_2 is considered in (11.54).

As in the previous models, some self-adjoint operators are preserved during the time evolution. These operators are $\hat{M}_j = \hat{S}_j + \hat{K}_j + \hat{I}_j + \hat{R}_j = \hat{\pi}_j + \hat{I}_j + \hat{R}_j$, $j = 1, 2$, where $\hat{R}_j = \int_{\mathbb{R}} r_j^\dagger(k) r_j(k)\, dk$. Then we can check that $[H, \hat{M}_j] = 0$, $j = 1, 2$. This implies that what is constant in time is the sum of the portfolio, the LoI and of the *overall reservoir input* of each trader. Notice that there is no general need, and in fact it is not required, for the cash or the number of shares to be constant in time. This is a measure of the fact that our SSM is not closed. This is completely different from what we assumed in Section 4.

The Heisenberg differential equations of motion can now be easily deduced:

$$\begin{cases} \frac{d}{dt} s_j(t) = -i\omega_j^s s_j(t) - i\lambda_{inf}\, i_j(t), \\ \frac{d}{dt} c_j(t) = -i\omega_j^c c_j(t) - i\lambda_{inf}\, i_j(t), \\ \frac{d}{dt} i_j(t) = -i\Omega_j i_j(t) - i\lambda_{inf}(s_j(t) + c_j(t)) - i\gamma_j \int_{\mathbb{R}} r_j(k,t)\, dk \\ \frac{d}{dt} r_j(k,t) = -i\Omega_j^{(r)}(k)\, r_j(k,t) - i\gamma_j\, i_j(t). \end{cases} \tag{11.55}$$

First of all, we rewrite the last equation in its integral form:

$$r_j(k,t) = r_j(k) e^{-i\Omega_j^{(r)}(k)t} - i\gamma_j \int_0^t i_j(t_1) e^{-i\Omega_j^{(r)}(k)(t-t_1)}\, dt_1,$$

and then we replace this in the differential equation for $i_j(t)$. Assuming that $\Omega_j^{(r)}(k) = \Omega_j^{(r)} k$, and following a somehow standard procedure, see [6], we deduce that

$$\frac{d}{dt} i_j(t) = -\left(i\Omega_j + \frac{\pi \gamma_j^2}{\Omega_j^{(r)}} \right) i_j(t) - i\gamma_j \int_{\mathbb{R}} r_j(k) e^{-i\Omega_j^{(r)} kt}\, dk - i\lambda_{inf}(s_j(t) + c_j(t)). \tag{11.56}$$

In the rest of this section we will work under the assumption that the last contribution in this equation can be neglected, when compared to the other ones. In other words, we are taking λ_{inf} to be very small. This procedure is slightly better than simply considering $\lambda_{inf}=0$ already in H, since we will keep the effects of this term in the first two equations in (11.55). Solving now (11.56) in its simplified expression, and replacing the solution $i_j(t)$ in the first equation in (11.55), we find:

$$s_j(t)=e^{-i\omega_j^s t}\left(s_j(0)-i\lambda_{inf}\alpha_j(t)\,i_j(0)-\lambda_{inf}\gamma_j\int_{\mathbb{R}}r_j(k)\,\eta_{2,j}(k,t)\,dk\right), \tag{11.57}$$

where we have defined

$$\alpha_j(t)=\frac{e^{(i\omega_j^s-\Gamma_j)t}-1}{i\omega_j^s-\Gamma_j},\qquad \eta_{2,j}(k,t)=\int_0^t\eta_{1,j}(k,t_1)e^{(i\omega_j^s-\Gamma_j)t_1}\,dt_1,$$

with

$$\Gamma_j=i\Omega_j+\frac{\pi\gamma_j^2}{\Omega_j^{(r)}},\qquad \eta_{1,j}(k,t)=\frac{e^{(\Gamma_j-i\Omega_j^{(r)}k)t}-1}{\Gamma_j-i\Omega_j^{(r)}k}.$$

It is clear from (11.55) that a completely analogous solution can be deduced for $c_j(t)$. The only difference is that ω_j^s should be replaced everywhere by ω_j^c.

The states of the system extend those of the previous section: for each operator of the form $X_{sm}\otimes Y_{res}$, where X_{sm} is an operator of the stock market and Y_{res} an operator of the reservoir, we have

$$\langle X_{sm}\otimes Y_{res}\rangle=\left\langle\varphi_{\mathcal{G}},X_{sm}\varphi_{\mathcal{G}}\right\rangle\omega_{res}(Y_{res}).$$

Here $\varphi_{\mathcal{G}}$ is of the form $\varphi_{\mathcal{G}}=\varphi_{S_1,K_1,I_1,S_2,K_2,I_2}$, in complete analogy with the other vectors considered all along this chapter, while $\omega_{res}(.)$ is a state satisfying again

$$\omega_{res}(1\!\!1_{res})=1,\quad \omega_{res}(r_j(k))=\omega_{res}(r_j^\dagger(k))=0,\quad \omega_{res}(r_j^\dagger(k)r_l(q))=N_j^{(r)}(k)\,\delta_{j,l}\delta(k-q),$$

for a suitable function $N_j^{(r)}(k)$. Also, $\omega_{res}(r_j(k)r_l(q))=0$, for all j and l. Then $N_{S_j}(t)=\left\langle s_j^\dagger(t)s_j(t)\right\rangle$ assumes the following expression:

$$N_{S_j}(t)=N_{S_j}(0)+\lambda_{inf}^2N_{I_j}(0)|\alpha_j(t)|^2+\lambda_{inf}^2\gamma_j^2\int_{\mathbb{R}}N_j^{(r)}(k)|\eta_{2,j}(k,t)|^2\,dk, \tag{11.58}$$

where $N_{I_j}(0)=\left\langle i_j^\dagger(0)i_j(0)\right\rangle=I_j$ and $N_{S_j}(0)=S_j$ are fixed by the quantum numbers of $\varphi_{\mathcal{G}}$. The expression for $N_{K_j}(t)=\left\langle c_j^\dagger(t)c_j(t)\right\rangle$ is completely analogous to the one above, with ω_j^s replaced by ω_j^c, and the portfolio of τ_j, $\pi_j(t)$, is simply the sum of $N_{S_j}(t)$ and $N_{K_j}(t)$. What we are interested in, is the variation of $\pi_j(t)$ over long time scales:

$$\delta\pi_j:=\lim_{t,\infty}\pi_j(t)-\pi_j(0).$$

Formula (11.58) shows that, if γ_j is small enough, the integral contribution is expected not to contribute much to $\delta\pi_j$, as this is proportional to γ_j^2. For this reason, we will not consider it in the rest of the section. We now find

$$\delta\pi_j = \lambda_{inf}^2 I_j(\Omega_j^{(r)})^2 \left(\frac{1}{\pi^2\gamma_j^4 + (\omega_j^s - \Omega_j)^2(\Omega_j^{(r)})^2} + \frac{1}{\pi^2\gamma_j^4 + (\omega_j^c - \Omega_j)^2(\Omega_j^{(r)})^2} \right). \quad (11.59)$$

Let us now recall that, at $t = 0$, the two traders are equivalent: $\omega_1^c = \omega_2^c =: \omega^c$, $\omega_1^s = \omega_2^s =: \omega^s$, $\Omega_1^{(r)} = \Omega_2^{(r)}$ and the initial conditions are $S_1 = S_2$, $K_1 = K_2$ and $I_1 = I_2$. The main difference between τ_1 and τ_2 is in Ω_1 which is taken larger than Ω_2: $\Omega_1 > \Omega_2{}^6$. With this in mind, we will consider three different cases: (a) $\gamma_1 = \gamma_2$; (b) $\gamma_1 > \gamma_2$; (c) $\gamma_1 < \gamma_2$. In other words, we are allowing a different interaction strength between the reservoir and the information term in H.

Let us consider the first situation (a): $\gamma_1 = \gamma_2$ and $\Omega_1 > \Omega_2$. In this case it is possible to check that $\delta\pi_1 < \delta\pi_2$, at least if $|\omega^c - \Omega_2| < |\omega^c - \Omega_1|$ and $|\omega^s - \Omega_2| < |\omega^s - \Omega_1|$. Notice that these inequalities are surely satisfied in our present assumptions if Ω_1 and Ω_2 are sufficiently larger than ω^c and ω^s. In this case the conclusion is, therefore, that the larger the LoI, the smaller the increment in the value of the portfolio. Needless to say, this is exactly what we expected to find in our model. Exactly the same conclusion is deduced in case (b): $\gamma_1 > \gamma_2$ and $\Omega_1 > \Omega_2$. In this case the two inequalities produce the same consequences: we are *doubling* the sources of the LoI (one from H_0 and one from the interaction), and this implies a smaller increment of π_1. Case (c): $\gamma_1 < \gamma_2$ and $\Omega_1 > \Omega_2$, is different. In this case, while H_0 implies that τ_1 is *less informed* (or that the quality of his information is not good enough), the inequality $\gamma_1 < \gamma_2$ would imply exactly the opposite. The conclusion is that, for fixed Ω_1 and Ω_2, there exists a critical value of (γ_1, γ_2) such that, instead of having $\delta\pi_1 < \delta\pi_2$, we will have exactly the opposite inequality, $\delta\pi_1 > \delta\pi_2$.

We should remind that these conclusions have been deduced under two simplifying assumptions which consist in neglecting the last contributions in (11.56) and in (11.58). Of course, to be more rigorous, we should also have some control on these approximations. However, we will not do this here.

As we see, this model looks completely reasonable and in fact was taken, in [16], as the starting point for a more complete situation, where an interaction between traders is also considered.

6 Conclusions

We have shown how operators can be used in the description of some macroscopic systems, giving rise to dynamical systems whose behavior is defined by a suitable operator, the Hamiltonian. We have discussed in some details how this Hamiltonian should be constructed and we have produced a set of *rules* which have been further used in two extremely different situations, i.e., in the description of a love affair and in the analysis of a SSM.

This line of research is going on following several different directions: from one side we are trying to construct other models for a SSM, in the attempt to approach some realistic version of stock market. From another side, we are using operators also in connection with decision making, and we have successfully applied our strategy to political alliances, just to cite the most recent application, [18]. Other macroscopic systems have also been considered in connection with population dynamics, [19, 20]. This short list of applications, which is not complete, shows how large is the range of applicability of our procedure.

What we believe would be important in this line of research is some simplified numerical approach: the unknown in our differential equations are in fact operators, and this increases significantly the difficulty of the procedure which gives some solution. This is a hard topic and is now work in progress.

Notes

1. This could be related also to psychological effects.
2. The existence of this limit has a long story which goes back, for instance, to [10] but which has been discussed by several authors along the years, see also [11] and references therein. We are not very interested in these mathematical aspects here. We just want to mention that this limit *creates problems*, but that these problems can be solved.
3. This will be assumed here: it is just a matter of choice of the parameters of the free Hamiltonian.
4. Of course, these differences must involve just two traders. Therefore, they must be related to just two values of j.
5. This number is not crucial, at least as far as H_{ex} is not considered. However, we make this choice to simplify the notation and our final deductions.
6. The case $\Omega_1 < \Omega_2$ can be easily deduced, by exchanging the role of Ω_1 and Ω_2.

Acknowledgments

The author acknowledges support from Università di Palermo and from GNFM. F.B. wishes to thank sincerely Prof. E. Haven for his contagious enormous scientific enthusiasm and for his support.

References

[1] C. Kittel, Introduction to solid state physics, John Wiley and Sons, New York, 1953.

[2] W. Segal, I. E. Segal, The Black–Scholes pricing formula in the quantum context, *Proc. Natl. Acad. Sci. USA*, **95**, 4072–4075, (1998).

[3] B.E. Baaquie, Quantum finance, Cambridge University Press, 2004.

[4] A. Khrennikov, Ubiquitous quantum structure: from psychology to finances, Springer, Berlin, 2010.

[5] E. Haven, A. Khrennikov, Quantum social science, Cambridge University Press, Cambridge, 2013.

[6] F. Bagarello, Quantum dynamics for classical systems: with applications of the number operator, John Wiley and Sons, New York, 2012.

[7] J. Busemeyer, P. Bruza, Quantum models of cognition and decision. Cambridge University Press 2012.

[8] F. Bagarello, F. Oliveri, An operator–like description of love affairs, *SIAM J. Appl. Math.*, **70**, 3235–3251, (2011).

[9] F. Bagarello, Damping in quantum love affairs, *Physica A*, **390**, 2803–2811 (2011).

[10] W. Thirring and A. Wehrl, On the Mathematical Structure of the B.C.S.-Model, *Commun. Math. Phys.* 4, 303–314, (1967).

[11] F. Bagarello, G. Morchio, Dynamics of mean field spin models from basic results in abstract differential equations, *J. Stat. Phys.* **66**, 849–866, (1992).

[12] F. Bagarello, An operatorial approach to stock markets, *J. Phys. A*, **39**, 6823–6840, (2006).

[13] Chao Zhang, Lu Huang, A quantum model for stock market, *Physica A*, in press.

[14] M. Schaden, Quantum finance, *Physica A*, **316**, 511–538, (2002).

[15] A. Messiah, Quantum mechanics, vol. 2, North Holland Publishing Company, Amsterdam, 1962.

[16] F. Bagarello, E. Haven, Towards a formalization of a two traders market with information exchange, *Phys. Scripta*, **90**, 015203 (2015).

[17] F. Bagarello, E. Haven, The role of information in a two-traders market, *Physica A*, **404**, 224–233, (2014).

[18] F. Bagarello, An operator view on alliances in politics, *SIAP*, **75**, (2), 564–584 (2015).

[19] F. Bagarello, F. Oliveri. An operator description of interactions between populations with applications to migration. *Math. Mod. Meth. Appl. Sci.*, **23**, 471–492, (2013).

[20] F. Bagarello, F. Gargano, F. Oliveri, A phenomenological operator description of dynamics of crowds: escape strategies, *Appl. Math. Model.*, doi:10.1016/j.apm.2014.10.038.

Index