ECONOMETRIC SOCIETY MONOGRAPHS

Advances in Economics and Econometrics

Theory and Applications, Eighth World Congress, Volume III

Edited by Mathias Dewatripont, Lars Peter Hansen, and Stephen J.Turnovsky

CAMBRIDGE

more information - www.cambridge.org/9780521818742

This page intentionally left blank

Advances in Economics and Econometrics

This is the third of three volumes containing edited versions of papers and commentaries presented at invited symposium sessions of the Eighth World Congress of the Econometric Society, held in Seattle, WA in August 2000. The papers summarize and interpret recent key developments, and they discuss future directions for a wide range of topics in economics and econometrics. The papers cover both theory and applications. Written by leading specialists in their fields, these volumes provide a unique survey of progress in the discipline.

Mathias Dewatripont is Professor of Economics at the Université Libre de Bruxelles where he was the founding Director of the European Centre for Advanced Research in Economics (ECARE). Since 1998, he has been Research Director of the London-based CEPR (Centre for Economic Policy Research) network. In 1998, he received the Francqui Prize, awarded each year to a Belgian scientist below the age of 50.

Lars Peter Hansen is Homer J. Livingston Distinguished Service Professor of Economics at the University of Chicago. He was a co-winner of the Frisch Prize Medal in 1984. He is also a member of the National Academy of Sciences.

Stephen J. Turnovsky is Castor Professor of Economics at the University of Washington and recently served as an Editor of the *Journal of Economic Dynamics and Control.* He is an Associate Editor and is on the Editorial Board of four other journals in economic theory and international economics.

Professors Dewatripont, Hansen, and Turnovsky are Fellows of the Econometric Society and were Program Co-Chairs of the Eighth World Congress of the Econometric Society, held in Seattle, WA in August 2000.

Econometric Society Monographs No. 37

Editors:
Andrew Chester, University College London
Matthew Jackson, California Institute of Technology

The Econometric Society is an international society for the advancement of economic theory in relation to statistics and mathematics. The Econometric Society Monograph Series is designed to promote the publication of original research contributions of high quality in mathematical economics and theoretical and applied econometrics.

Other titles in the series:
G. S. Maddala *Limited dependent and qualitative variables in econometrics*, 0 521 33825 5
Gerard Debreu *Mathematical economics: Twenty papers of Gerard Debreu*, 0 521 33561 2
Jean-Michel Grandmont *Money and value: A reconsideration of classical and neoclassical monetary economics*, 0 521 31364 3
Franklin M. Fisher *Disequilibrium foundations of equilibrium economics*, 0 521 37856 7
Andreu Mas-Colell *The theory of general economic equilibrium: A differentiable approach*, 0 521 26514 2, 0 521 38870 8
Truman F. Bewley, Editor *Advances in econometrics – Fifth World Congress (Volume I)*, 0 521 46726 8
Truman F. Bewley, Editor *Advances in econometrics – Fifth World Congress (Volume II)*, 0 521 46725 X
Hervé Moulin *Axioms of cooperative decision making*, 0 521 36055 2, 0 521 42458 5
L. G. Godfrey *Misspecification tests in econometrics: The Lagrange multiplier principle and other approaches*, 0 521 42459 3
Tony Lancaster *The econometric analysis of transition data*, 0 521 43789 X
Alvin E. Roth and Marilda A. Oliviera Sotomayor, Editors *Two-sided matching: A study in game-theoretic modeling and analysis*, 0 521 43788 1
Wolfgang Härdle, *Applied nonparametric regression*, 0 521 42950 1
Jean-Jacques Laffont, Editor *Advances in economic theory – Sixth World Congress (Volume I)*, 0 521 48459 6
Jean-Jacques Laffont, Editor *Advances in economic theory – Sixth World Congress (Volume II)*, 0 521 48460 X
Halbert White *Estimation, inference and specification*, 0 521 25280 6, 0 521 57446 3
Christopher Sims, Editor *Advances in econometrics – Sixth World Congress (Volume I)*, 0 521 56610 X
Christopher Sims, Editor *Advances in econometrics – Sixth World Congress (Volume II)*, 0 521 56609 6
Roger Guesnerie *A contribution to the pure theory of taxation*, 0 521 23689 4, 0 521 62956 X
David M. Kreps and Kenneth F. Wallis, Editors *Advances in economics and econometrics – Seventh World Congress (Volume I)*, 0 521 58011 0, 0 521 58983 5
David M. Kreps and Kenneth F. Wallis, Editors *Advances in economics and econometrics – Seventh World Congress (Volume II)*, 0 521 58012 9, 0 521 58982 7
David M. Kreps and Kenneth F. Wallis, Editors *Advances in economics and econometrics – Seventh World Congress (Volume III),* 0 521 58013 7, 0 521 58981 9
Donald P. Jacobs, Ehud Kalai, and Morton I. Kamien, Editors *Frontiers of research in economic theory: The Nancy L. Schwartz Memorial Lectures, 1983–1997*, 0 521 63222 6, 0 521 63538 1
A. Colin, Cameron and Pravin K. Trivedi, *Regression analysis of count data*, 0 521 63201 3, 0 521 63567 5
Steinar Strøm, Editor *Econometrics and economic theory in the 20th century: The Ragnar Frisch Centennial Symposium*, 0 521 63323 0, 0 521 63365 6
Eric Ghysels, Norman R. Swanson, and Mark Watson, Editors *Essays in econometrics: Collected papers of Clive W. J. Granger (Volume I)*, 0 521 77297 4, 0 521 80401 8, 0 521 77496 9, 0 521 79697 0
Eric Ghysels, Norman R. Swanson, and Mark Watson, Editors *Essays in econometrics: Collected papers of Clive W. J. Granger (Volume II)*, 0 521 79207 X, 0 521 80401 8, 0 521 79649 0, 0 521 79697 0
Cheng Hsiao, *Analysis of panel data*, second edition, 0 521 81855 9, 0 521 52271 4
Mathias Dewatripont, Lars Peter Hansen, and Stephen J. Turnovsky, Editors *Advances in economics and econometrics – Eighth World Congress (Volume 1)*, 0 521 81872 8, 0 521 52411 3
Mathias Dewatripont, Lars Peter Hansen, and Stephen J. Turnovsky, Editors *Advances in economics and econometrics – Eighth World Congress (Volume II)*, 0 521 81873 7, 0 521 52412 1

Advances in Economics and Econometrics

Theory and Applications, Eighth World Congress, Volume III

Edited by

Mathias Dewatripont
Université Libre de Bruxelles and CEPR, London

Lars Peter Hansen
University of Chicago

Stephen J. Turnovsky
University of Washington

CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE UNIVERSITY PRESS
Cambridge, New York, Melbourne, Madrid, Cape Town, Singapore, São Paulo

Cambridge University Press
The Edinburgh Building, Cambridge CB2 2RU, United Kingdom

Published in the United States of America by Cambridge University Press, New York

www.cambridge.org

Information on this title: www.cambridge.org/9780521818742

First published in print format 2003

ISBN-13 978-0-511-06991-8 eBook (EBL)
ISBN-10 0-511-06991-X eBook (EBL)

ISBN-13 978-0-521-81874-2 hardback
ISBN-10 0-521-81874-5 hardback

ISBN-13 978-0-521-52413-1 paperback
ISBN-10 0-521-52413-X paperback

Contents

Contributors

Stefania Albanesi
Bocconi University and IGIER

Orazio P. Attanasio
University College London, IFS, NBER, and CEPR

Abhijit V. Banerjee
Massachusetts Institute of Technology

V. V. Chari
University of Minnesota and Federal Reserve Bank of Minneapolis

Lawrence J. Christiano
Northwestern University and Federal Reserve Banks of Chicago and Cleveland

Francis X. Diebold
University of Pennsylvania and NBER

Jordi Galí
CREI and Universitat Pompeu Fabra

Kenneth L. Judd
Stanford University

Felix Kubler
Stanford University

Lucrezia Reichlin
Université Libre de Bruxelles (ECARES) and CEPR

José-Víctor Ríos-Rull
University of Pennsylvania, NBER, and CEPR

Karl Schmedders
Northwestern University

Christopher A. Sims
Princeton University

Mark W. Watson
Princeton University

Preface

These volumes contain the papers of the invited symposium sessions of the Eighth World Congress of the Econometric Society. The meetings were held at the University of Washington, Seattle, in August 2000; we served as Program Co-Chairs. Volume 1 also contains an invited address, the "Seattle Lecture," given by Eric Maskin. This address was in addition to other named lectures that are typically published in *Econometrica*. Symposium sessions had discussants, and about half of these wrote up their comments for publication. These remarks are included in the book after the session papers they comment on.

The book chapters explore and interpret recent developments in a variety of areas in economics and econometrics. Although we chose topics and authors to represent the broad interests of members of the Econometric Society, the selected areas were not meant to be exhaustive. We deliberately included some new active areas of research not covered in recent Congresses. For many chapters, we encouraged collaboration among experts in an area. Moreover, some sessions were designed to span the econometrics–theory separation that is sometimes evident in the Econometric Society. We followed the lead of our immediate predecessors, David Kreps and Ken Wallis, by including all of the contributions in a single book edited by the three of us. Because of the number of contributions, we have divided the book into three volumes; the topics are grouped in a manner that seemed appropriate to us.

We believe that the Eighth World Congress of the Econometric Society was very successful, and we hope that these books serve as suitable mementos of that event. We are grateful to the members of our Program Committee for their dedication and advice, and to Scott Parris at Cambridge University Press for his guidance and support during the preparation of these volumes. We also acknowledge support from the officers of the Society – Presidents Robert Lucas, Jean Tirole, Robert Wilson, Elhanan Helpman, and Avinash Dixit – as well as the Treasurer, Robert Gordon, and Secretary, Julie Gordon. Finally, we express our gratitude to the Co-Chairs of the Local Organizing Committee, Jacques Lawarree and Fahad Khalil, for a smoothly run operation.

CHAPTER 1

Contracting Constraints, Credit Markets, and Economic Development

Abhijit V. Banerjee

1. INTRODUCTION

Development economists are, perhaps by necessity, optimistic people. One does not become a development economist if one believes that the world's poorest are doing as well as they possibly could. Indeed, the premise of the entire field[1] is that there is talent in every people, if not every person, and if there is one central question, it has to be: What prevents people from making the best use of their natural talents?

There are at least five distinct answers to this question. The first, which is elaborated here, is the answer from contract theory: Talent is not an apple[2]; one cannot simply go to the market, sell one's talent, and expect to be paid the appropriate price. The second is coordination failure: Talent is talent only if it gets to work with the appropriate other inputs. Even Lennon needed Paul and George – had they decided to go to the City instead, he too might have found himself a different profession.[3] The third is political economy: Governments can and often do make it harder for people to do what they are best at doing.[4] The fourth is learning: People may not know what they ought to be doing, and even when they do the rest of the world may not appreciate them. For example, a growing body of evidence shows that farmers are often ignorant or suspicious

[1] And of growth theory: What is convergence other than the hope that there is talent in every nation?

[2] A contract theorist would probably say an apple is not an apple either – an apple can be stale or fresh, sweet, or sour. If you had a lot of apples to sell, you would probably want to invest in a reputation for selling only fresh and sweet apples.

[3] There is a long tradition in development economics of models that emphasize coordination failures going back at least to Rosenstein-Rodan (1943) and Nurkse (1953). See Murphy, Shleifer, and Vishny (1989) for a model where there is failure of coordination between producers of different goods, and Kremer (1993) for a model of matching between different types of talents.

[4] There is, of course, a long tradition here, going back to Adam Smith. Krueger (1974) and Bhagwati (1982), among others, study the distortions in the allocation of talent and resources that come from government policies.

of more rewarding crops and better seeds.[5] The final answer comes from what has come to be called behavioral economics: People may not always seek out the best options because they are held back by psychological constraints or social norms.[6]

The fact that this survey concentrates on the contract theoretic argument should not be interpreted as evidence for its primacy. However, it is the argument that has received the most elaboration over the past decade or so, and the one that best matches the competencies of the present author. It is therefore the appropriate topic for a survey such as this.

2. CONTRACT THEORY IN DEVELOPMENT ECONOMICS

Contract theoretic arguments in development economics go back at least to the work of D. G. Johnson in the 1940s and 1950s in the context of land markets.[7] Stiglitz's 1974 paper on sharecropping, among others, started a tradition of formal contract theoretic models that seek to explain why landlords and tenants often settle into arrangements that are, at least apparently, less than first best efficient.[8]

Since then, similar principles have been applied to the study of all the other important markets: capital, insurance, and human capital. The result is an enormous literature that I could not even begin to do justice to within the limits of this survey. I confine myself, therefore, to elaborating on a single example from the market for capital, which I hope will allow me to draw out the most important themes, though at several points in the text I point out the connections with what is understood about the other asset markets.

3. THE CREDIT MARKET

The facts about the credit market are remarkably stark. Although neoclassical theory predicts a single price of capital at which people both borrow and lend, at any point over the past twenty years one could point to a set of peoples in the world (most recently the Japanese) who were earning a negative return on their savings, while another set of people were borrowing at real rates of 60 percent or more.

[5] See Besley and Case (1994) and Munshi (2000). Banerjee (1992) provides theoretical arguments for why such behavior may be rational for individual farmers.

[6] There is a long and controversial literature on this point. The famous Lewis model (Lewis, 1954) argued that family norms could discourage people from seeking outside options. The rational peasant model (Schultz, 1964) was articulated as a critique of models like the Lewis model (see also Cole, Mailath, and Postlewaite, 1992, and Banerjee and Newman, 1998, for two very different attempts to reconcile these views). It is, however, time to revisit this issue: With the increasing sophistication of the psychological models used in economics, it is now possible to reask the question of whether, for example, poverty can have direct discouragement effects.

[7] See Johnson (1950).

[8] See Stigltiz (1974) and Cheung (1968).

Indeed, more often than not, very large differences between borrowing and lending rates can be found within a single subeconomy. Banerjee (2001) reviews a number of empirical studies of individual credit markets in developing countries and lists six salient features.[9]

First, there is a sizeable gap between lending rates and deposit rates within the same subeconomy. Ghatak (1976) reports data on interest rates paid by cultivators in India from the *All India Rural Credit Survey* for the 1951–1952 to 1961–1962 period: The average rate varies between a maximum of 18 percent (in 1959–1960) and a minimum of about 15 percent (in 1961–1962). These numbers are, however, slightly misleading: Around 25 percent of the borrowing reported in these surveys was zero-interest loans, usually from family members or friends. These should be seen as gifts or insurance rather than loans. If these were left out, the average rates in these surveys would be above 20 percent. We are not told what the comparable rates for depositors were in this period, but Ghatak reports that the bond rate in this period was around 3 percent, and the bank deposit rate was probably about the same.

Timberg and Aiyar (1984) report data on indigenous style bankers in India, based on surveys that they carried out. They report that the gap between the average rate charged to borrowers and the average rate to depositors by finance companies was 16.5 percent. The same gap for financiers from the Shikarpuri community was 16.5 percent, 12 percent for financiers from the Gujerati community, 15.5 percent for the Chettiars, 11.5 percent for the Rastogis, and so on.

The "Summary Report on Informal Credit Markets in India" (Dasgupta, 1989) reports results from a number of case studies that were commissioned by the Asian Development Bank and carried out under the aegis of the National Institute of Public Finance and Policy. For the rural sector, the data are based on surveys of six villages in Kerala and Tamil Nadu, carried out by the Centre for Development Studies. The average interest rate charged by professional moneylenders (who provide 45.61 percent of the credit) in these surveys is about 52 percent. Although the average deposit rate is not reported, the maximum from all the case studies is 24 percent and the maximum in four out of the eight case studies is no more than 14 percent. For the urban sector, the data are based on various case surveys of specific classes of informal lenders: For finance corporations, they report that the maximum deposit rate for loans of less than a year is 12 percent, whereas the minimum lending rate is 48 percent. For hire-purchase companies in Delhi, the deposit rate was 14 percent and the lending rate was at least 28 percent. For autofinanciers in Namakkal, the gap between the deposit rate and the lending rate was 19 percent.[10]

[9] The review focuses on the informal sector because the formal banking sector in most developing countries has tended to be quite rigid (interest rate caps, strict rules about collateral, inflexible credit limits, etc.; see Ghate, 1992), with the result that the informal sector has become the supplier of the marginal units of capital for all but the very largest of firms.

[10] This number and all other information about this gap are measured in percentage points.

For handloom financiers in Bangalore and Karur, the gap between the deposit rate and the lowest lending rate was 26 percent.[11]

Aleem (1990) reports data from a study of professional moneylenders that he carried out in a semiurban setting in Pakistan in 1980–1981. The average interest rate charged by these lenders is 78.5 percent. The bank rate in that year in Pakistan was 10 percent. However, it is possible that depositors in this area may not have been depositing in the banks, so an alternative measure of the gap can be obtained by using Aleem's numbers for the opportunity cost of capital to these moneylenders, which is 32.5 percent.[12]

Second, there is extreme variability in the interest rate charged by lenders for superficially similar loan transactions within the same economy. Timberg and Aiyar (1984) report that the rates for Shikarpuri financiers varied between 21 percent and 37 percent on loans to members of local Shikarpuri associations and between 21 percent and 120 percent on loans to nonmembers (25 percent of the loans were to nonmembers and another 50 percent were loans through brokers). In contrast, the Gujerati bankers charged rates of no more than 18 percent. Moreover, the rates faced by established commodity traders in the Calcutta and Bombay markets were never above 18 percent and could be as low as 9 percent.

The "Summary Report on Informal Credit Markets in India" (Dasgupta, 1989) reports that finance corporations offer advances for a year or less at rates between 48 percent per year and the utterly astronomical rate of 5 percent per day. The rates on loans of more than a year varied between 24 percent and 48 percent. Hire-purchase contracts offer rates between 28 percent and 41 percent per year. Handloom financiers charge rates between 44 percent and 68 percent, yet the Shroffs of Western India offer loans at less than 21 percent and Chit Fund members can borrow at less than 25 percent.

The same report tells us that, among rural lenders, the average rate for professional moneylenders (who in this sample give about 75 percent of the commercial informal loans) was 51.86 percent, whereas the rate for the agricultural moneylenders (farmers who also lend money) who supply the rest was 29.45 percent. Within the category of professional moneylenders, about half the loans were at rates of 60 percent or more but another 40 percent or so had rates below 36 percent.

The study by Aleem (1990) reports that the standard deviation of the interest rate was 38.14 percent compared with an average lending rate of 78.5 percent. In other words, an interest rate of 2 percent and an interest rate of 150 percent are both within two standard deviations of the mean.

Swaminathan (1991) reports on a survey of two villages in South India that she carried out: The average rate of interest in one village varied between

[11] A number of other lending institutions are also mentioned in this study. However, the range of both deposit rates and lending rates is so wide in these cases that the gap between the minimum lending rate and the maximum deposit rate is not very large. This does not rule out the possibility that the gap between the *average* borrowing and lending rate is quite substantial even in these cases.

[12] This, however, understates the gap, because the moneylenders themselves borrow this money, and the original lenders are paid much less than 32.5 percent.

14.8 percent for loans collateralized by immovable assets (land, etc.) and 60 percent for loans backed by movable assets. The corresponding rates in the other village were 21 percent and 70.6 percent. Even among loans collateralized by the same asset – gold – the average rate in one village was 21.8 percent but it went up to 58.8 percent when the loans were to landless laborers.

Ghate (1992) reports on a number of case studies from all over Asia: The case study from Thailand found that interest rates were 2–3 percent per month in the Central Plain but 5–7 percent in the North and Northeast (note that 5 percent and 7 percent are very different).

Gill and Singh (1997) report on a survey of six Punjab villages that they carried out. The mean interest rate for loans up to Rs. 10,000 is 35.81 percent for landowning households in their sample, but 80.57 percent for landless laborers.

Fafchamps' (2000) study of informal trade credit in Kenya and Zimbabwe reports an average monthly interest rate of 2.5 percent (corresponding to an annualized rate of 34 percent) but also notes that this is the rate for the dominant trading group (Indians in Kenya, whites in Zimbabwe). Blacks pay 5 percent per month in both places.[13]

Irfan et al. (1999) report that interest rates charged by professional moneylenders vary between 48 percent and 120 percent.

Third, there are low levels of default. Timberg and Aiyar (1984) report that average default losses for the informal lenders they studied range between 0.5 percent and 1.5 percent of working funds.

The "Summary Report on Informal Credit Markets in India" (Dasgupta, 1989) attempts to decompose the observed interest rates into their various components,[14] and it finds that the default costs explain 14 percent (not 14 percentage points!) of the total interest costs for the Shroffs, around 7 percent for autofinanciers in Namakkal and handloom financiers in Bangalore and Karur, 4 percent for finance companies, 3 percent for hire-purchase companies, and essentially nothing for the Nidhis. The same study reports that, in four case studies of moneylenders in rural India, default rates explained about 23 percent of the observed interest rate.

The study by Aleem gives default rates for each individual lender. The median default rate is between 1.5 percent and 2 percent and the maximum is 10 percent.

Fourth, production and trade finance are the main reasons given for borrowing, even in cases where the rate of interest is relatively high. Ghatak (1976) concludes on the basis of his study that "the existing belief about the unproductive use of loans by Indian cultivators . . . has not been substantiated."

Timberg and Aiyar (1984) report that, for Shikarpuri bankers (who charge 31.5 percent on average, and as much as 120 percent on occasion), at least 75 percent of the money goes to finance trade and, to a lesser extent, industry.

[13] Fafchamps notes that, when he controls for the sector of the economy, and so on, this difference goes away, but that just tells us that the source of the variation is sector rather than race.

[14] In the tradition of Bottomley (1963).

The "Summary Report on Informal Credit Markets in India" (Dasgupta, 1989) reports that several of the categories of lenders that have been already mentioned, such as hire-purchase financiers (interest rates between 28 percent and 41 percent), handloom financiers (44–68 percent), Shroffs (18–21 percent), and finance corporations (24–48 percent for longer-term loans and more than 48 percent on loans of less than a year), focus almost exclusively on financing trade and industry, and even for Chit Funds and Nidhis, which do finance consumption, trade and industry dominate.

Swaminathan (1991) reports that, in the two villages she surveys, the share of production loans in the portfolio of lenders is 48.5 percent and 62.8 percent. The higher share of production loans is in Gokalipuram, which has the higher interest rates (above 36 percent for all except the richest group of borrowers). Ghate (1992) also concludes that the bulk of informal credit goes to finance trade and production.

Murshid (1992) studies Dhaner Upore loans in Bangladesh (you get some amount in rice now and repay some amount in rice later) and argues that most loans in his sample are production loans despite the fact that the interest rate is 40 percent for a three- to five-month loan period.

Gill and Singh (1997) report that the bulk (63.03 percent) of borrowing from the informal sector goes to finance production. This proportion is lower for the landless laborers, but it is a nonnegligible fraction (36 percent).

Fifth, richer people borrow more and pay lower rates of interest. Ghatak (1976) correlates asset category with borrowing and debt in the *All India Rural Credit Survey* data and finds a strong positive relationship. Timberg and Aiyar (1984) report that some of the Shikarpuri and Rastogi lenders set a credit limit that is proportional to the borrower's net worth: Several lenders said that they would lend no more than 25 percent of the borrower's net worth, although another said he would lend up to 33 percent.

The "Summary Report on Informal Credit Markets in India" (Dasgupta, 1989) tells us that, in its rural sample, landless laborers paid much higher rates (ranging from 28 percent to 125 percent) than cultivators (who paid between 21 percent and 40 percent). Moreover, Table 15.9 in that report clearly shows that the average interest rate declines with loan size (from a maximum of 44 percent to a minimum of 24 percent). The relation between asset category and interest rate paid is less clear in their data, but it remains that the second poorest group (those with assets in the range Rs. 5,000–10,000) pay the highest average rate (120 percent) and the richest (those with more than Rs. 100,000) pay the lowest rate (24 percent).

Swaminathan (1991) finds a strong negative relation between the value of the borrower's land assets and the interest rate he or she faces: The poorest (those with no land assets) pay 44.9 percent in one village and 45.4 percent in the other, whereas the rich (those with land valued at more than Rs. 50,000) pay 16.9 percent and 24.2 percent in the corresponding villages.

Gill and Singh (1997) show that the correlation between the wealth of the borrower and loan size is negative after the interest rate is controlled for. They also find a positive relation between the borrower's wealth and the loan he or she gets.

Sixth, bigger loans are associated with lower interest rates. Table 15.9 in the "Summary Report on Informal Credit Markets in India" (Dasgupta, 1989) clearly shows that the average interest rate declines with loan size (from a maximum of 44 percent to a minimum of 24 percent).

Ghate (1992) notes that the interest rate on very small loans in Bangladesh tends to be very high (Taka 10 per week on a loan of Taka 500, or 86 percent per annum).

Gill and Singh (1997) show that the correlation between loan size and the interest rate is negative even after they control for the wealth of the borrower.

3.1. Taking Stock: The Facts About Credit Markets

The fact that there is a gap between the lending rate and the rate paid to depositors is not, per se, surprising. The fact that intermediation is costly is, after all, entirely commonplace. What is striking is the size of the gap. It is always more than 10 percent and usually more than 14 percent, in a world where interest rates paid to depositors are rarely more than 20 percent and usually closer to 10 percent. In other words, intermediation costs seem to eat up at least a third and often half (and sometimes much more than half) of the income that could go to depositors.

However, this argument overstates the point slightly. The probability that a moneylender would default on his or her deposit liabilities is substantially lower than the probability that borrowers would default on the loan, which implies that the default premium on loans should be much greater than the default premium on deposits. From the evidence reported herein, default is relatively rare and default costs rarely raise the interest rate by more than 10 percent. The gap between loan rates and deposit rates would be very large even if we were to deduct 10 percent from the loan rate.[15, 16]

The fact that interest rates vary quite so much is particularly striking given the standard neoclassical prediction that in market equilibrium the marginal unit of capital in every firm should earn the same return. However, given that people might be rationed in the credit market, it is theoretically possible that the marginal product of capital is actually equal in all its uses, despite the enormous disparities in the interest rate. Note that the incremental capital/output ratio for the Indian economy is estimated to be around 4.3, implying a marginal return

[15] One might be worried that although default rates are low on average, default may be very important in those cases where the interest rate is high. However, this is not a problem because, for the most part, we look at interest rates and default rates weighted by volume (or equivalently, do a Bottomley, 1963, decomposition). Moreover, in the one detailed microstudy we have in which the average interest rate is very high (Aleem, 1990), default rates are actually very low (always less than 10 percent, and usually less than 2 percent).

[16] Delay in repayment for which no extra interest is charged is another factor that raises the interest rate. In Aleem's data, delay is much more common than default, but in a significant fraction of the cases the lender is able to charge interest for the extra days. Moreover, the percentage of loans that are late never exceeds 25 percent and the average delay is no more than six months, so at worst this would raise the interest rate by a factor of 1.12.

on capital of 24 percent. This is, however, a gross measure and the true return, net of depreciation, is clearly substantially lower (no more than 20 percent). The fact that interest rates above 35 percent are standard, and those above 75 percent are by no means rare, suggests that at least some of the users of capital must value capital at substantially more than 20 percent.

Could it be that all of the demand at relatively high interest rates comes from people who have particularly insistent consumption needs today? This is certainly not the stated purpose of the loans, as already noted. Of course, money is fungible and one cannot rule out the possibility that some of these people are either deluded or untruthful. However, it remains that when a handloom producer borrows at 48 percent or more to finance consumption, he or she chose to do so instead of taking the money out of his or her existing business. Therefore, it must be that the handloom producer would lose more than 48 percent on any money that comes out of his or her business. This may be in part because, in the short run, his or her assets are not liquid, but this could not explain why the producer accepts ongoing financing on these terms. Therefore, the producer must be earning marginal returns that are close to 48 percent.[17]

The fact that the marginal product of capital varies substantially across borrowers in the same subeconomy is supported by more direct evidence from the knitted garment industry based on data that I collected in joint work with Kaivan Munshi (Banerjee and Munshi, 2001). Tirupur produces 70 percent of India's knitted garment exports, and India is a major exporter of knitted garments. There are two communities of producers in Tirupur: Gounders, who are linked by community ties to a rich local agricultural community; and Outsiders, a motley crew of businessmen from all over India. They produce exactly the same goods, yet they use radically different technologies. Gounders invest much more than Outsiders at all levels of experience, both in absolute terms and relative to output. Average capital/output ratios for Gounders can be three times as large as that for Outsiders and is typically twice as large. However, all the evidence points to the Outsiders being more able: They enjoy faster output growth, and their output outstrips that of the Gounders after a few years.

One possible situation in which high-ability people may invest less would be if capital were less useful for them, which would be the case if ability and capital were substitutes in the production function. The evidence, however, points against this explanation: When we compare Gounders with Gounders or Outsiders with Outsiders, it is clear that those who grow faster and produce more also invest more. Therefore, it seems relatively clear that the Outsiders invest less *despite having a higher marginal product of capital.*[18]

[17] There is, once again, the possibility that a part of the reason why these rates are so high is because of default risk. In other words, the expected return on the marginal units of capital need not be as high as 48 percent. However, as already observed, defaults contribute relatively little to the level of the interest rate.

[18] This is what Caballero and Hammour (2000) call scrambling. The proximate reason, it appears, is the Gounders have a lot of investiable funds that they cannot profitably lend out because

What explains why the credit markets behave in this way? Why is intermediation so inefficient with some people and so efficient with others? Why are the rich borrowers and those who borrow more favored by the market?

The standard theory of interest rates decomposes them into default rates, opportunity cost, transaction costs, and monopoly rents. This is useful descriptively but stops well short of an explanation – the problem is that none of these can be seen as independent causal factors. Take the example of default rates. The fact that default rates are relatively low is not a fact about the nature of default: Timberg and Aiyar (1984) observe that some branches of the state-owned commercial banks in India have default rates up to 60–70 percent. The low default rates observed in the studies we mention are a result of the steps taken by lenders to avoid default.

Monitoring the borrower is an obvious example of the kind of steps that lenders take. It is also an important source of what goes under the rubric of transaction costs: Aleem (1990) and Irfan et al. (1999) provide a list of steps taken by the lender to avoid default. These include getting to know the borrower through other transactions, visiting the borrower's establishment, making inquiries about the borrower, and going after the borrower to make sure he or she repays.

Lenders also protect themselves by limiting their lending to borrowers they know.[19] This has four important consequences. First, it pushes capital toward well-connected borrowers and away from less well-connected borrowers, even when there is no difference in their productivity. Second, it makes it important that lending be local – the lender must know and trust his or her borrowers. This adds one or more layers of intermediation to the process of lending, with additional transaction costs entering at each of the stages, which raises the opportunity cost of capital. Third, it forces the lender to limit his or her lending, with the consequence that both the lender's capital and skills as a lender may remain unused for a significant part of the time. This raises both the opportunity cost of the capital and the transaction cost (which includes a part of the lender's time). Finally, it gives the lender some ex post monopoly power, as a borrower would find it hard to leave a lender who knows him or her well. Under competitive conditions, these ex post rents will be dissipated in ex ante competition, with lenders in effect subsidizing new borrowers in order to extract rents later from those who will become his or her long-term clients.

What this tells us is that the four components of the interest rate are all jointly determined in the process of the lender making his or her lending decisions. Depending on the lender's strategy, it could be that the transaction costs dominate

intermediation is so inefficient in India. Instead, they set up their own garment firms or lend to friends and family in the garment business. Because these firms are set up as a conduit for this surplus capital, they are not required to be particularly productive. The Outsiders, by contrast, come from traditional entrepreneurial communities and, as a result, their capital probably has many alternative uses. In other words, they do not invest in Tirupur because they lack other choices. This makes them more likely to be productive but also less willing to invest a lot.

[19] See McMillan and Woodruff (1999).

or the opportunity cost dominates, or that default or monopoly rents become very important. The strategy could be very different depending on the nature of the clientele and other environmental characteristics. This may be a part of the reason why different people have taken very different views of informal credit markets: Aleem, for example, finds that for every rupee lent, about half a rupee goes into transaction costs, whereas Dasgupta (1989) finds that only about 30 percent of interest costs are explained by transaction costs (strictly establishment costs); Ghate (1992) argues that transaction costs are unimportant except in the case of very small loans.[20]

The fact that all these decisions are interrelated clearly makes it dangerous to use any single one of these components as a measure of the efficiency of intermediation. For example, Ghate (1992) sees the low level of transaction costs in his sample as evidence for the remarkable efficiency of informal lending. But, as has already been noted, transaction costs may be low because the lenders are very choosy about to whom they lend. This raises the opportunity cost of capital (because capital is often idle) and limits credit access, both of which have their welfare costs. Likewise, the low rate of defaults in informal transactions is often mentioned as evidence for their efficiency, but this is obviously misleading if it comes at the cost of increased monitoring or reduced lending. Finally, the presence of rents in lending is not, per se, evidence for lack of competition in the market. As pointed out herein, in this type of market, ex post rents are consistent with ex ante competition.

A further implication of this observation is that both loan size and the interest rate are jointly determined, and therefore one cannot give a causal interpretation of the relation between interest rates and loan size reported herein. Rather, one should see both of these as outcomes that are determined by more primitive variables, such as the wealth of the borrower, the borrower's productivity, the liquidity of his or her assets, and so on. This also makes it harder to interpret the reported negative relation between the borrower's wealth and the interest rate. In principle, it could be entirely a result of the fact that rich borrowers borrow more.

What is most important is that this line of argument underscores the significance of developing a proper theory of credit markets. Such a theory would explain the variation in interest rates and the gap between interest rates and deposit rates in terms of the true primitives of the model, and make predictions about the relation between loan size and interest rates and borrower and lender characteristics. Although there is a long tradition of models of imperfect credit markets going back to Jaffee and Russell (1976) and Stiglitz and Weiss (1981), and the arguments behind why credit markets can fail, based on moral hazard and/or adverse selection, are well known, I feel that it is useful to develop a framework that has a more direct empirical orientation.

[20] Even within the set of case-studies reported by Ghate, there seems to be considerable variation. In Kerala, the case-study concludes that transaction costs are of negligible importance while the Thai study concludes that transaction costs added between 3 and 14 percentage points to the interest cost.

3.2. A Simple Model of Moral Hazard in the Credit Market

There is an investment opportunity whose gross returns are $F(K)R(p)$ with probability p and 0 otherwise, where K is the amount invested and $F(\cdot)$ is a production function. If an investor wants to invest more than his or her wealth, W, the investor will need to borrow. There is a capital market and the (gross) cost of capital in that market is ρ. To make this problem interesting, we assume the following.

1. p is a choice for the investor but is unobserved by the lender. p takes a value between p_0 and p_1.
2. $E(p) \equiv pR(p)$ has the property that $E'(p_0) > 0$, and $E''(p) \leq 0$.
3. The only possible contract is a loan contract.[21]

3.2.1. The Basic Moral Hazard Problem

The optimal value of p, p^*, is clearly greater than p_0 and may or may not be less than p_1. The combination of the rest of the assumptions tells us that there is no guarantee that p^* would be chosen in equilibrium. To see this, note that the borrower, who is assumed to be risk neutral, will choose p to maximize $F(K)E(p) - pr(K - W)$, where r is the interest rate that has to be paid to the lender to make him or her willing to lend.

The borrower will choose p such that $E'(p)F(K) - r(K - W) = 0$.[22] This is quite obviously inconsistent with the social optimum: the borrower clearly wants to choose $p < p^*$. This is the standard incentive problem in credit markets: Society cares about net output but the borrower cares only about what remains after paying interest. This is the essence of all models of ex ante moral hazard in the credit market.

Next, notice that the first-order condition for the borrower's choice of p can be rewritten in the following form:

$$E'(p)\frac{F(K)}{K} = r\left(1 - \frac{W}{K}\right). \tag{3.1}$$

From this equation it is evident that p depends on three things: the average product of capital, $F(K)/K$; the leverage ratio, K/W; and the interest rate, r. If capital is more productive, the borrower is less inclined to misbehave, and this is reflected in a lower p. Being more leveraged worsens the borrower's incentives and so does a higher interest rate, which is consistent with the observation made earlier that the interest cost burden is the source of the distortion.

[21] This rules out making the borrower's payments depend on the project's realized returns. Diamond (1989) justifies this assumption by assuming that the realized return is not publicly observable except by making use of a liquidation proceeding, which is costly to the point of using up all available output. This makes sure that a borrower will not willfully default as long as the lender threatens to go into liquidation whenever he or she defaults.

[22] Assuming an interior optimum exists.

Property 1 (Efficiency). There is less inefficiency in the credit relationship when there is less leveraging, when the interest rate is lower, and when the project is more productive.

From this it follows that the equilibrium value of p can be written in the following form:

$$p = p(R, F(K)/K),$$

where $R = r[1 - (W/K)]$ is the interest cost per unit of investment. Clearly $\partial p/\partial R < 0$, and $(\partial p)/[\partial F(K)/K] > 0$. Writing the relation in this form draws attention to the important role played by the shape of the production function. When $F(\cdot)$ is concave, $F(K)/K$ decreases as a function of K. Therefore, those who invest more will be more liable to moral hazard, even after controlling for the leverage ratio. However, if F is convex, at least over a range, increasing the level of investment may increase profitability and improve the borrower's incentives. As we will see, this distinction may be very important for some questions.

3.2.2. The Interest Rate

We have so far treated r as a parameter. In fact, if there is competition in lending, lenders should not make any profits, which would imply that

$$r = \rho/p, \tag{3.2}$$

or

$$R = \frac{\rho(1 - W/K)}{p} = \frac{\Gamma}{p},$$

where Γ is the cost of capital per unit of investment.[23] Solving $p = \underset{\sim}{p}(R, F(K)/K)$ along with $R = \Gamma/p$ gives us $p = \tilde{p}(\Gamma, F(K)/K)$ and $R = \tilde{R}(\Gamma, F(K)/K)$. However, it is easy to construct examples where these equations have multiple solutions: Intuitively, a fall in p raises r, but a rise in r, as we already saw, puts downward pressure on p. It is not clear, however, that we can interpret these as multiple equilibria – if the lender knows the rules of the game, then the lender knows that he or she can pick the best equilibrium and make everyone better off, simply by setting the right interest rate. Therefore, unless

[23] The assumption of perfect competition in the credit market is not uncontroversial. There is a long tradition of papers that view high interest rates as evidence for monopoly power in the credit market. However, as already pointed out, the issue of rents in the credit market is likely to be quite delicate, because competition operates ex ante rather than ex post. Therefore, the absence of ex ante rents is consistent with Bhaduri's (1977) model of how lenders trap borrowers into a permanent cycle of debt and debt repayment. The evidence seems to support the hypothesis of ex ante competition: The few studies (Ghate, 1992, Dasgupta, 1989, and Aleem, 1990) that compute the gap between the interest rate charged and the various costs of lending (opportunity cost, monitoring costs, and default costs) do not find a large gap on average, though one cannot reject the possibility that there is a large rent component in many individual transactions.

the lender is boundedly rational, we should probably assume that the best equilibrium is always chosen. This is the equilibrium with the lowest interest rate.

Assuming that this is the equilibrium, the comparative statics of the $p(\cdot)$ function are inherited by the $\widetilde{p}(\cdot)$ function, and $\widetilde{r} = \rho/\widetilde{p}$ shares the properties of the $\widetilde{p}$ function, only reversed. A lower leverage ratio increases p and lowers the interest rate, as does a higher average product of capital. Lowering the cost of capital lowers the rate of interest more than proportionately because the repayment rate goes up.

Property 2 (Interest rates). Borrowers who are more leveraged tend to pay higher rates, whereas more productive borrowers pay lower rates. Raising the cost of capital raises the interest rate more than proportionately.

3.2.3. The Level of Investment

The next step is to endogenize the level of borrowing. The borrower's choice of K maximizes

$$F(K)E(p) - \rho(K - W),$$

under the assumption that p depends on K through the $\widetilde{p}(\cdot)$ function. The first-order condition for that maximization is

$$F'(K)E(p) + F(K)E'(p)\frac{\partial p}{\partial F(K)/K}\frac{\partial F(K)/K}{\partial K} + F(K)E'(p)\frac{\partial p}{\partial \Gamma}\frac{\rho}{W} = \rho. \tag{3.3}$$

If we compare this with the first-order condition in a first best world, $F'(K)E(p^*) = \rho$, we see that there are three sources of distortion. First, $E(p) < E(p^*)$, which says that capital is less productive and therefore the borrower wants to invest less. Second, $\partial p/\partial \Gamma$ is negative, which also discourages investment. Finally, there is the second term on the left-hand side, which can be positive or negative depending on the sign of $[\partial F(K)/K]/(\partial K)$. This, as we have already observed, depends on whether the production function is concave or not. If it is concave, the second term is negative and it is unambiguously true that imperfections in the capital market lead to less investment. If not, the second term may be positive, and if this effect is large enough, it could outweigh the first effect and generate overinvestment. Whether this possibility is actually worth taking seriously remains an open question, awaiting more precise calibrations of the model.[24]

Another important property of the first best is that the amount invested is independent of the wealth of the investor. In our present model, if we were to increase W, keeping K fixed, we know from Property 1 that p would go up, raising $E(p)$ and reducing $E'(p)$. As long as F is concave, both of these effects go in the same direction: They both raise the rewards for investing more,

[24] Lehnert, Ligon, and Townsend (1999) argue that this is a real possibility.

and therefore there is more investment.[25] In fact, in the special case where $F(K) = \sigma K$, that is, a linear production technology, K not only goes up when W goes up, it is precisely proportional to W.

The general case of a nonconcave F tends to be complex. One interesting example, in which there is a single indivisible investment, turns out to be very straightforward. In this case people either invest or do not, and because those who have more wealth choose a higher p at the same level of investment (Property 1), they are the ones who will invest. More generally, nonconvex production technologies raise the possibility that the poor will actually invest more than the rich: Intuitively, if the production function is convex, increasing investment raises productivity, which improves incentives through its direct effect. However, there is also an indirect effect: Investing more makes the borrower more leveraged and this worsens incentives. The balance of these two effects may be different for the rich and the poor, because their incentive problems are different, and in principle it could be that the poor end up investing more. However, it seems unlikely that these effects would dominate the main effect of being richer, which is that (at the same level of investment) richer people are less leveraged and therefore have better incentives and as a result their capital is more productive.

Lowering the cost of capital in this model increases p, and this, as argued herein, encourages investment. However, lowering the cost of capital also increases the amount invested in the first best, so that there is no clear prediction for the extent of underinvestment.

Property 3 (The level of investment). Capital market imperfections lead to underinvestment in the typical case, though it is not inconceivable that they could generate overinvestment. The more wealthy will tend to invest more in absolute terms. When the production technology is linear, the amount invested will be proportional to the investor's wealth. When there is a single indivisible investment, the rich are more likely to invest than the poor. Lowering the cost of capital increases investment.

Capital market imperfections reduce the demand for capital. For a given supply curve of capital, this means that the cost of capital will be lower than it would be otherwise. In the longer run, however, the supply of capital will also respond to the pattern of wealth creation generated by the capital market imperfection, and the net impact on the cost of capital is ambiguous.

Property 4 (The cost of capital). For a given supply curve for capital, imperfect capital markets will have a lower cost of capital, but this is no longer necessarily true once we take into account the impact of the capital market imperfection on the supply of credit.

[25] Actually, there is a third effect: Increasing W/K, it can be shown, reduces $[\partial p/\partial F(K)]/K$, thereby reinforcing the effect of the fall in $E'(p)$.

3.2.4. Introducing Monitoring

The model developed so far is useful in developing intuition about how the credit market works but it has an important limitation in terms of explaining the data. As we have already seen, the repayment rates in most informal credit transactions are very high (over 90 percent). It follows from Equation (3.2) that the interest charged by a competitive lender can be only about 10 percent higher than the cost of capital, which from all the evidence given herein is much too small a margin.

The missing piece of the story is monitoring. We have assumed so far that the lender cannot do anything to affect the borrower's choice of p. This is clearly an extreme assumption, because, as already mentioned, lenders can and do monitor borrowers.

The point of all these activities is to learn more about the borrower. This helps in two ways: first, by allowing the lenders to pick borrowers for whom the interval $[p_0, p_1]$ is relatively small, thereby limiting the possibility of moral hazard, and second, by getting to know the borrower's environment, thereby making it easier to find out when the borrower is not doing what he or she has promised to do with the money.

In addition to this kind of ex ante monitoring, there is ex post monitoring of the project, which is checking that the borrower has done what he or she had promised to do with the money. For example, the lender can try to make sure that the borrower is spending the money on inputs for his or her project rather than on consumption. Finally, there is collection: Once the loan falls due, the lender has to spend time chasing each overdue loan.

It is not possible to capture all of these different aspects of monitoring in a single model, so the discussion here is limited to one specific model, though some of the other models are discussed in a later section. We introduce monitoring into the model by assuming that if the lender monitors at a level a, the borrower will chose a project $p(a)$ or a project with a p no lower than $p(a)$.[26] We assume that this comes about through either ex ante monitoring of the project (screening of projects before the loan is given) or ex post monitoring of the project (checking on the borrower after he or she has been given the loan and punishing the borrower if he or she has not done what he or she was supposed to do). The problem is that we know very little about the nature of the empirical relation between monitoring and project choice. The only option we have is to reason on purely a priori grounds.

One assumption that has a certain plausibility is that the amount of monitoring necessary is a function of the extent of misalignment of incentives between the borrower and the lender. The borrower in our model wants to choose $p = p(R, F(K)/K)$, which gives him or her a payoff of $F(K)E(p(R, F(K)/K)) - p(K/W, r, F(K)/K)r(K - W)$, whereas the lender wants the borrower to choose p, which gives him or her a payoff of $F(K)E(p) - pr(K - W)$. The

[26] The borrower will typically choose the lowest permissible value.

extent of misalignment is therefore

$$D = F(K)[E(p(R, F(K)/K)) - E(p)] - [p(R, F(K)/K) - p]RK. \tag{3.4}$$

Our assumption is then that the amount of monitoring is a function of D. However, to allow for different types of scale effects, we write it in a slightly more general form:

$$M = M(K, D/K, m),$$

where m is a parameter that shifts the monitoring cost function, $dM/dm > 0$.

3.2.5. The Cost of Capital With Monitoring

The lender's participation constraint (3.2) now takes the form

$$R = \frac{\Gamma}{p} + \frac{M(K, D/K, m)}{Kp}. \tag{3.5}$$

This equation defines $R(\Gamma, K, p, m)$, the interest rate for a borrower with a fixed W who wants to invest an amount K and promises to choose a project p. Using this, we can define the expected cost of credit per unit of investment: $C(\Gamma, K, p, m) = pR$.

This formulation of the supply side of credit has the obvious advantage that the interest rate can be much higher than the cost of capital even if defaults are rare. This is because monitoring costs can be very high; indeed the reason why there is very little default may be a result of the resources spent on monitoring.

It is useful to begin our analysis of this model with an examination of the properties of the $C(\cdot)$ function. Simple differentiation tells us that

$$\frac{\partial R}{\partial \Gamma} = \frac{1}{p - \{p - p[R, F(K)/K]\}/K \ \partial M/[\partial(D/K)]},$$

$$\frac{\partial R}{\partial m} = \frac{\partial M/\partial m}{K(p - \{p - p[R, F(K)/K]\}/K \ \partial M/[\partial(D/K)])}.$$

Because $(p - \{p - p[R, F(K)/K)]\}/K\partial M/[\partial(D/K)]) < p$, this tells us that increases in the cost of lending (represented by a rise in ρ or in m) have a multiplier effect, resulting in a bigger increase in the interest rate than would be warranted by the direct effect of the increase in cost.[27] This is because the initial rise in the interest rate worsens the borrower's incentives and makes it necessary that the borrower be monitored more, which raises the cost of lending even further, and so on. This property is obviously also inherited by the $C(\cdot)$ function.

[27] In principle, this increase can be very large because $(1 - \{p - p[R, F(K)/K]\}/K\partial M/[\partial(D/K)])$ can be very close to zero or even negative (in which case, the equilibrium interest changes discontinuously).

Property 5 (Multiplier). The interest rate and the amount of monitoring can be very sensitive to changes in the cost of capital and/or the cost of monitoring.

This is an important property: It tells us that there are even relatively small differences in the monitoring cost, or the cost of capital can induce a lot of variation in the interest rate, which helps to explain why we observe so much variation.

A related and important property of the $C(\cdot)$ function comes from differentiating equation (3.5) with respect to p. This, after some algebraic manipulations, gives us

$$\frac{\partial C}{\partial p} = \frac{1}{K^2}\frac{\partial M}{\partial (D/K)}\frac{p(R, F(K)/K)RK - pF(K)E'(p)}{p - \{p - p[R, F(K)/K]\}/K\ \partial M/[\partial (D/K)]}.$$

Equation (3.1) tells us that $E'(p(R, F(K)/K))F(K) = RK$. Using this (assuming that $p - (1/K)\partial M/[\partial (D/K)](p - p(R, F(K)/K)) > 0$), we find it immediately clear that the sign of $\partial C/\partial p$ depends on the sign of $p(R, F(K)/K)E'(p(R, F(K)/K)) - pE'(p)$. Because $p > p(R, F(K)/K)$, it follows that C_p can be positive only if the function $pE'(p)$ is a decreasing function of p over a range.

This makes it clear that it is entirely possible that C_p be negative for all p, implying that implementing high values of p may, paradoxically, require less monitoring than implementing lower values. This is because a high p generates a low R and this improves incentives.

In such situations it may be optimal to raise p all the way to its maximum, that is, to p_1. In particular, this will be true as long as $E(p)$ is everywhere increasing in p over its admissible range,[28] and it will remain true *irrespective of how costly it is to monitor*.[29] However, it is easy to see that it will never be optimal for p to exceed its social welfare maximizing level, that is, the value of p for which $E'(p) = 0$. This is because when $E'(p) = 0$, C_p is clearly positive.

Property 6 (Default). Very low levels of default may be optimal even when monitoring is quite costly, though it is never optimal to have less default than in the social optimum.

This is important because it tells us that it is often optimal to aim for very low default rates even at the cost of lots of costly monitoring and high interest rates. This is reassuring, given that the combination of very low default rates and very high interest rates is by no means uncommon.[30]

[28] For example, C_p is negative whenever $E(p)$ takes the form Ap^β, with $A > 0$ and $\beta \in (0, 1)$.

[29] Of course, this is conditional on the loan contract being viable, which is not the case when monitoring is too costly.

[30] Aleem's data set from Pakistan, mentioned herein, is an example.

3.2.6. The Optimal Credit Contract With Monitoring

The optimal credit contract will be a combination (K, p) that maximizes

$$F(K)E(p) - pR(\Gamma, K, p, m)KM.$$

The first-order conditions that describe the optimal contract (when it is not a corner solution) are

$$F(K)E'(p) = KC_p, \tag{3.6}$$

$$F'(K)E(p) = C + KC_K. \tag{3.7}$$

There is, however, relatively little that we can say about the optimal credit contract at this level of generality. The problem is easily seen from Equation (3.5): An increase in K affects both the numerator and the denominator of the expression $[M(K, D/K, m)]/[(K - W)p]$, and without more structure it is not possible to say anything about how more investment affects the expected cost of lending.

The Model With Constant Returns in Monitoring. One simple and fruitful way to impose structure is to assume constant returns in monitoring; that is,

$$M(K, D/K, m) = KM(D/K, m).$$

For the most part, we will also assume that there are constant returns in production; that is, $F(K) = \sigma K$. In this case, Equation (3.5) can be rewritten in the following form:

$$pR = \Gamma + M(\sigma[E(p(R, \sigma) - E(p)] - [p(R, \sigma) - p]R, m).$$

R is therefore a function of σ, m, p, and Γ, and so is, therefore, the expected cost of lending C. It follows that, keeping p and Γ fixed, doubling the borrower's wealth and the amount he or she invests does not change the unit cost of lending. It follows that all borrowers with the same σ and the same m will choose the same leverage ratio and face the same interest rate. In other words, under full constant returns, the rich and the poor will pay the same rate of interest as long as they are equally productive. The rich will simply invest more.

The direct prediction of this model is the absence of a correlation between the borrower's wealth and the interest rate. Of course, as we will see later, it does not rule out the possibility of a spurious correlation, induced by a correlation between W and either σ or m. Nevertheless, this result provides a useful benchmark: It tells us that there is no necessary reason why the rich should pay lower interest rates, as observed in the data.

The Model With a Fixed Cost of Monitoring. A model that manages to account for most of the observed fairly economically is one in which there is a fixed cost of monitoring that has to be paid as long as there is some borrowing and a variable cost, which, as before, exhibits constant returns:

$$M(K, D/K, m) = KM(D/K, m) + \Phi.$$

In this case, Equation (3.5) can be rewritten to read

$$pR = \Gamma + \Phi/K + M(\sigma[E(p(R,\sigma) - E(p)] - [p(R,\sigma) - p]R, m). \tag{3.8}$$

With these assumptions, we run the risk that the lender's maximization problem may not be convex: To see why, note that $\Gamma + \Phi/K = \rho(1 - W/K) + \Phi/K$, which tells us that if $\Phi > \rho W$, R goes down when K goes up, encouraging the borrower to borrow even more. Conversely, a borrower who borrows little will pay very high rates, making it attractive for this borrower not to borrow at all, suggesting the possibility of a "bang-bang" solution. As long as we make sure that σ is not too large (to avoid the possibility that the demand for credit becomes infinite), the solution will be for the poor borrower ($\rho W \ll \Phi$) to borrow nothing.

However, there is an interior solution for borrowers who are richer ($\rho W \gg \Phi$). It is easily checked that this interior solution has a very simple property: From Equation (3.8) it follows that as long as p, σ, ρ, and m are held fixed, R is completely determined by the term $(\rho W - \Phi)/K$ and therefore $C \equiv C[(\rho W - \Phi)/K]$. From Equation (3.7) the optimal choice of K satisfies

$$\sigma = C\left(\frac{\rho W - \Phi}{K}\right) + \frac{\rho W - \Phi}{K} C'\left(\frac{\rho W - \Phi}{K}\right)$$

in this case, which tells us that $(\rho W - \Phi)/K$ is uniquely determined by σ. A number of properties follow immediately from this observation. First, an increase in W results in a more than proportional increase in K; in other words, richer people are more leveraged. Second, higher values of Φ are associated with a lower value of K. Third, changes in W and Φ do not affect R, from which it follows that r goes down when W goes up (because K/W goes up and R remains unchanged).

There is a straightforward intuition behind these results. Given that there is a fixed cost of lending, those who invest more will face a lower cost of capital. However, for a poor person to be able to invest the same amount as a richer person, the leverage ratio would have to be much higher, and this distorts incentives. The optimal contract balances these two types of costs: The poor end up both investing less and paying more in interest.

As a way to assess whether the model generates the right orders of magnitude, the model was simulated under the assumption that $E(p) = 2p^{0.5}$, and $M(D/K, m) = mD/K$.[31] For parameter values $\rho = 1.05$, $m = 0.8$, $\Phi = 0.5$, and $\sigma = 0.66$, we find that those with wealth levels up to about 1.75 (i.e., about three and a half times the fixed cost of monitoring) do not invest at all. When investment begins, the interest rate is above 50 percent (and the leverage ratio

[31] Note that $pE'(p)$ is increasing in p so that it is always optimal to choose the highest possible value of p. We set this value to be 0.9, so that the default rate is fixed at 10 percent (which is high but within the observed range).

is 2.8), and as the borrower's wealth goes up, the interest rate goes down and converges to about 27 percent, while the leverage ratio rises and converges to about 4.2.

Property 7 (Wealth effects). When there are constant returns in both production and monitoring, two borrowers who differ only in their wealth levels will be equally leveraged and will pay the same interest rate. When there is a fixed cost of monitoring, richer borrowers will pay a lower rate and will be more leveraged, and the very poorest borrowers will prefer not to borrow at all.

These wealth effects have the implication that the wealth advantage the rich start with will tend to get amplified: first, because the difference in the amount invested is typically going to be larger than the difference in wealth. This follows from the fact that the leverage ratio is greater than one and either constant or increasing in wealth.[32] Second, as long as R is increasing in K, the marginal product of capital will be higher than the expected interest cost in equilibrium and each unit of investment generates some pure profits; this follows from Equation (3.7). Because the rich invest more than the poor, they earn more pure profits.

3.3. Taking Stock: How to Think About Credit Markets

The model with a fixed cost of monitoring gives us a simple way of accounting for the facts that are listed herein. Moreover, it helps to explain the fact that short-run interest rates in informal markets are often higher than longer-run interest rates (see, e.g., Table 3.2 in Dasgupta, 1989). Ghate (1992) also notes that very short short-term loans are often particularly expensive. On the face of it, this is puzzling because one would imagine that the scope for moral hazard is greater in longer-term contracts. The fixed cost approach can resolve this puzzle as long as there is a part of the fixed cost that is transaction specific and independent of the length of the contract.

However, not surprisingly, this is not the only way to account for these facts. For example, if the production function were concave rather than linear, the rich would be less leveraged than the poor (because diminishing returns set in at high levels of investment). As a result, the interest rate they face will tend to be lower. Although they are less leveraged, the absolute amount that the rich borrow may still be higher. High levels of credit will therefore be associated with lower interest rates. Similar patterns may arise if, for example, it is cheaper to monitor the rich because, say, the rich share closer social ties with those who are lending.

In any case, there is no reason why we should confine ourselves to models of ex ante moral hazard. Ex post moral hazard (borrowers who try to avoid repaying) is clearly an important aspect of credit markets and so is adverse

[32] Note, however, that this property may not hold if all the wealth was not liquid. In that case, the leverage ratio may be less than one.

selection.[33] There are also other ways to model ex ante moral hazard: Holmstrom and Tirole (1996) developed a model where the borrower wants to put in less effort rather than take too much risk. Many of the basic predictions of our model show up in these models as well. Not surprisingly, less leveraged borrowers, all else being the same, tend to be better credit risks and face lower interest rates at the same level of borrowing. Consequently, richer borrowers will have their initial advantage compounded by the workings of the credit market. High interest rates, as before, make the borrower more likely to misbehave in these models as well.[34] This, in turn, tends to raise the interest rate (either because default becomes more likely or because more monitoring is called for). Therefore, what I call the multiplier property of interest rates – namely the fact that small increases in the costs of lending can lead to a large increase in the interest rate – ought to be true in these other models as well. Finally, although the cost of monitoring is rarely formally introduced into these models (Holmstrom and Tirole, 1996, being an important exception), it is intuitively clear that the shape of the monitoring cost function will play a crucial role and a fixed cost of monitoring will have an effect similar to the one discussed herein.

One might even want to venture beyond models where the borrower is the main source of moral hazard. The paper by Holmstrom and Tirole, mentioned herein, worries about the incentives of intermediaries in the credit market. They argue that these problems translate into a further credit constraint, this time at the level of the intermediary. Moreover, the typical intermediary is large, and if the intermediary itself has the right incentives, its agents who do the actual business of lending – the loan officers and credit appraisers of the world – may not, because what they have at stake personally is only a very small part of the amount of money they control. The solution is typically to restrict the domain they control: Stringent bureaucratic rules about what loan officers can and cannot do are a feature of credit market intermediaries the world over. This, of course, comes at a cost. Credit decisions become bureaucratized and typically much less responsive to all but the hardest information. Assessments of the quality of the project and judgments about future profitability, both relatively soft information, will tend to have little impact on who gets credit. To the extent that other institutions, such as venture capitalists, do not pick up the slack, this will hurt new entrants and the most radical ideas. This problem may also be most serious where the banking sector is dominated by the public sector, given that there is already a tendency toward bureaucratization and buck passing.[35]

[33] Hart and Moore (1994) provided the best-known model of ex post moral hazard in the credit market; Stiglitz and Weiss (1981) provided the classic reference on adverse selection.

[34] Though, as pointed out in Aghion, Banerjee, and Piketty (1999), in models of ex post moral hazard, there is a possible countervailing effect coming from the fact that high interest rates make it more credible for the lender to put a lot of effort into pursuing recalcitrant borrowers.

[35] For evidence that this is a very real problem, see Banerjee and Duflo (2001). There are also many anecdotes that support this view: It is said, for example, that Indian bankers in the 1980s and early 1990s were puzzled by how they could justify lending to software companies, because their only real assets were their staff, and work in progess consisted of lines of code on the computer.

Of course, much more work (simulations, etc.) is called for before we can be sure that these alternative theories can generate the right orders of magnitude. But in one sense there is no reason to pose these as alternatives. All of them working together can generate a larger aggregate effect, and large aggregate effects are clearly important in giving relevance to this class of theories. However, from the point of view of actually designing policy, it is important to know exactly where the constraint lies. It is also, from the point of view of both macro relevance and micro policy design, important to identify the exact structure of the credit constraint. Is the amount of credit that a borrower can get primarily a function of the borrower's wealth and his or her expected profitability, as our model suggests, or is it the case that profitability is largely ignored, as the simple bureaucratic model sketched herein would suggest? How important is the distinction, ignored in our analysis, between wealth and liquid wealth – one additional reason why the poor may suffer is that their wealth may be less liquid?[36] Is the borrower's wealth the key ingredient or is it his or her inventories, as in the classic models of inventory financing?[37] The answer to each of these questions has many important ramifications, and careful empirical research on the technology of lending remains one of the imperatives of the day.

Theme 1. The observed patterns in credit markets – low default rates, high and highly variable interest rates, and credit limits that increase with wealth – suggest that contracting in credit markets is highly imperfect and monitoring is very important. This suggests that there will be underinvestment, and a significant part of the output produced will be wasted on monitoring. The earnings gap between the rich and the poor will be amplified by the capital market imperfection, and this will be particularly the case if, as seems plausible, there are fixed costs of monitoring.

4. THE DYNAMICS OF WEALTH ACCUMULATION

One important message from the previous section is that the poor are at a disadvantage in the credit market. However, as we have seen, the exact form of the disadvantage tends to depend on the technologies of production and monitoring. For example, in the case in which there are constant returns in both production and monitoring, the disadvantage takes the form of a proportional reduction in the amount they invest with no difference in the interest rate or the

[36] The work by Swaminathan, already cited, suggests that this may indeed be an important distinction. The requirement that the wealth be held as liquid collateral also creates a demand for collateralizable assets, and shifts in the relative price of these assets become important (see Kiyotaki and Moore, 1997, for a macroeconomic model based on this particular relative price shift).

[37] Based presumably on the idea that loans have to be fully collateralized and inventories can serve as collateral. Clearly, the plausibility of this model depends on how easy it is to attach inventories.

choice of projects, whereas a model with a fixed cost of monitoring generates variations in the interest rate and in the choice of projects. In the short run, however, all versions of the capital market imperfection have the common implication that the poor will be underrewarded for their talent.

The longer-run implications of the different models are, however, potentially very different. To see this, imagine a world where there is one good produced and a population of identical people who each live for one period and always have one child. Each person starts life with an endowment that her parent gave her. Her life is simple, verging on the drab. At the beginning of her life, she chooses among income earning opportunities. Her choices will be either to invest in a productive opportunity or to lend out the money. The exact technology of production will be discussed later, but it could be thought of as either investing in learning a skill, starting a business, or even patenting or promoting a new idea. At the end of the period, the person decides on what to do with her realized income, which consists of her investment earnings plus an endowment e – she can leave it to her child or eat it herself. For simplicity, assume that she has Cobb–Douglas-like preferences over consumption (c) and bequest (b):

$$U(c, b) = A[c^{1-\beta} b^{\beta}], \quad 0 < \beta < 1, \ A > 0.$$

Because the person allocates her end-of-period wealth between these two uses, this immediately implies that, if her end-of-period income (or wealth) is y,

$$c = (1 - \beta)y, \quad b = \beta y.$$

People borrow to invest more than what they were born with. Assume that the credit market is exactly as modeled in the previous section.

These assumptions together define a simple dynamic process that maps W_t, the wealth of an individual from the current generation (which is equal to the bequest he received), into the wealth of his child, W_{t+1}. The exact shape of this map will depend on what we assume about technology, which is what we turn to now.

First consider the case in which both the production technology and the monitoring technology are fully linear. In this case we know that the optimal leverage ratio, the optimal level of monitoring, and the optimal choice of p are all independent of the investor's wealth. To simplify life, assume that the optimal choice of p is 1. Denote the monitoring cost required to sustain this by m (per dollar lent). Finally, assume that the production technology takes the form $F(K) = E(1)K$ and that the optimal leverage ratio is λ. In this case, the map from current wealth to future wealth is given by

$$W_{t+1} = \beta[e + \kappa W_t],$$

where $\kappa = E(1)\lambda - (\lambda - 1)(\rho + m)$.

In the case in which $\beta\kappa < 1$, this is a process that converges to a wealth level $\beta e/(1 - \beta\kappa)$: Every dynasty ends up with the same wealth in the long run.

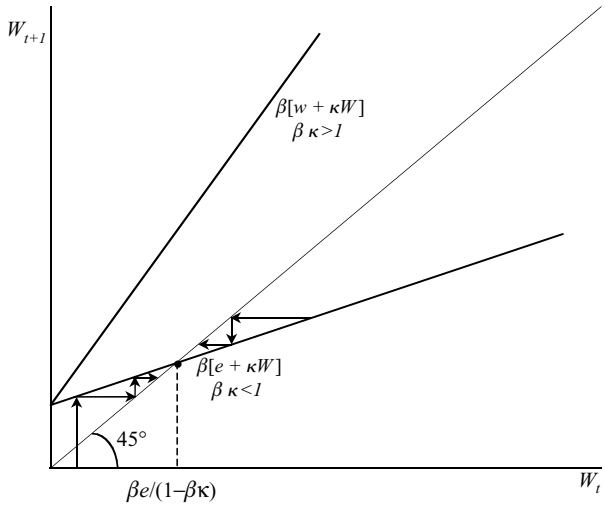


Figure 1.1.

In contrast, in the case in which $\beta\kappa \geq 1$, no dynasty ever converges but in the long run the wealth of every dynasty grows at the same rate and everyone becomes extremely wealthy. There are no poverty traps in this model.[38] The dynamics of wealth in this model are shown in Figure 1.1.

The model behaves in much the same way when the production and monitoring technologies are strictly convex (i.e., both functions are strictly concave). A simple way to see this is to note that the poor always have the option of choosing the same leverage ratio as the rich. If they were to choose the same leverage ratio, they would actually pay less in interest than the rich, because they have a higher marginal product of capital and monitoring them is easier (diminishing returns in monitoring). Their net return per dollar of their wealth would therefore be higher, and as a result, the wealth of the poor will grow faster than that of the rich.[39]

Things change significantly when at least one of the technologies stops being convex. To take an example, assume that the production technology remains what it was in the previous example but introduce a fixed cost of monitoring. Under these assumptions, the equation representing the evolution of wealth for

[38] However, this model is consistent with divergence across countries: Those countries that have a better financial system, and therefore a lower m, will grow faster.

[39] This claim does depend in an important way on the fact that there is only one type of investment. Mookherjee and Ray (2000) show that, if there are many alternative types of investment and the economy needs all of them, long-run inequality may be inevitable in a world of imperfect capital markets, even if all the individual technologies are convex and people are forward looking in their savings decisions.

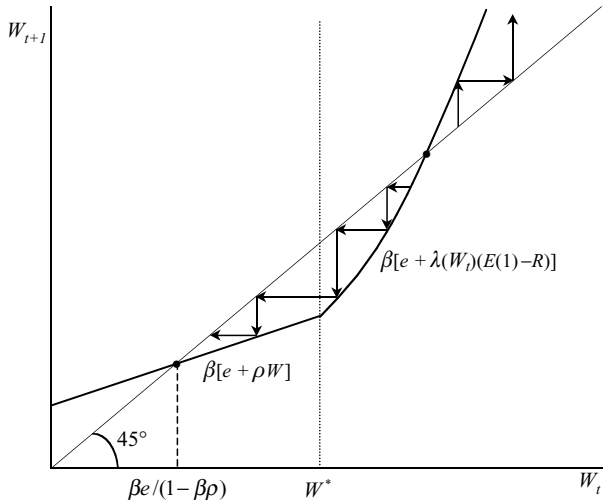


Figure 1.2.

those who can borrow can be written in the following form:

$$\begin{aligned} W_{t+1} &= \beta[e + E(1)K_t - (K_t - W_t)r] \\ &= \beta e + \beta\lambda(W_t)(E(1) - R). \end{aligned}$$

As shown in the previous section, when there is a fixed cost of monitoring, R is constant but λ is an increasing function of wealth and converges to $\overline{\lambda}$ as wealth continues to go up. For wealth below some W^*, people do not borrow and their wealth evolves according to

$$W_{t+1} = \beta e + \beta W_t E(1).$$

It is easy to see that this generates a map that has the convex shape shown in Figure 1.2.

The wealth dynamics implied by this picture are very different from those generated by the linear model. The rich get richer all the time, but the poor converge to a steady state at $\widetilde{W}$. However, it is not neccessary that inequality grows without a bound: It is straightforward to add to this model a technological limit on investment, $\overline{K}$. It is easily checked that what this does is to reduce the amount borrowed by those who would have otherwise invested more than $\overline{K}$. Eventually, when someone's wealth exceeds $\overline{K}$, he or she will stop borrowing and start lending, which causes the W_{t+1} schedule to take the form

$$W_{t+1} = \beta[e + E(1)\overline{K} + (W_t - \overline{K})\rho].$$

As long as $\beta\rho < 1$, this implies that wealth will remain bounded: By the fact that the curve is continuous, it follows that it must be S shaped. Depending

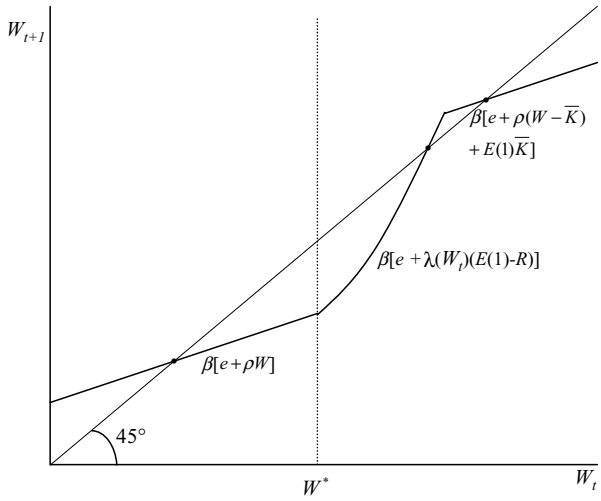


Figure 1.3.

on the parameters, this model has either one or three steady states, of which the two extreme ones are stable. When there are two stable steady states, as in Figure 1.3, the poor converge to the lower one and the rich to the upper, reflecting the fact, already noted, that when there is a fixed cost of monitoring, the poor earn a lower return on their wealth.

This is the classic poverty trap situation: The poor remain poor and the rich stay rich. Versions of this story have been told in many papers, including those by Galor and Zeira (1993), Dasgupta and Ray (1986), and Banerjee and Newman (1994), though they each have a different name for the investment: education in Galor and Zeira; health in Dasgupta and Ray; and capital in Banerjee and Newman. Both Galor and Zeira and Dasgupta and Ray have explicit nonconvexities in the production function, whereas Banerjee and Newman introduce the nonconvexity into the technology of lending. Moav (1999) presents a slightly different version of the argument, which relies on nonconvexities in what one might call the bequest technology.[40]

Of course, there are good reasons not to take poverty traps literally. The very lucky and the very talented among the poor will probably manage to escape their background, and some of the rich will surely manage to squander their patrimony. The robust implication of this model is rather that economic mobility

[40] In terms of our notation, the assumption he makes is that the amount bequeathed is a strictly convex function of end-of-period income over some range, whereas we have so far assumed that it is a linear function. It is easy to see that this can generate a poverty trap; essentially the rich benefit from the fact that rich parents leave bequests that are disproportionately large relative to their wealth.

will be slow. An economy that starts with many poor people will remain both poor and unequal for a very long time. This, however, raises the question of whether, in the end, this model is very different from the model where all the technologies are convex and there is no poverty trap, but the capital markets are extremely inefficient – after all, when capital markets do not work very well, convergence can take a very long time.

My view is that it is nevertheless worth distinguishing between these two models because the forms of mobility that they permit are rather different. The convex model predicts a slow and steady rise for all the poor, which culminates in their catching up with the rich (or if there is no convergence, they still become very, very rich). The mobility in the model with nonconvexities, by contrast, comes from those who are either very talented or very lucky. In other words, it takes the form of large jumps by a relatively few people.[41] This difference also shows up in the shape of the long-run distributions. In the convex model, most people in the long run will be middle class, with some outliers who are either very lucky or very unlucky. In contrast, in the other model most people will be either rich or poor.

Theme 2. Models in which all the key technologies – the production technology, the monitoring technology, and the bequest technology – are all linear or convex tend to favor long-run convergence: Those of comparable talent will earn comparable amounts in the long-run and the long-run distribution of income will reflect the distribution of abilities in the population. In contrast, models in which at least one of these technologies is nonconvex can generate poverty traps at the level of the individual. People who start poor stay poor, and therefore equally able people may earn very different amounts even in the long run.

4.1. Endogenous Savings and Poverty Traps

The one obviously unsatisfactory aspect of our model so far is the modeling of bequest decisions. Although the way we have modeled them is extremely convenient and there is no good reason, either empirical or a priori, to switch to full "Barro" preferences, there are clearly cases in which our model seems a bit strained. In the specific example in the previous section where monitoring has a fixed cost, the rate of return on beginning-of-period wealth (i.e., bequests) varies enormously and those who start with more will, over a range, earn a higher return. It is therefore plausible that parents will take this into account when planning their bequests.

To see what changes when we allow for endogenous savings decisions, consider a modification of our basic model that makes the agents infinitely

[41] Paulson and Townsend (2000) and Jeong and Townsend (2000) make a related but different distinction between models (such as Evans and Jovanovic, 1989) in which poor capital markets hurt the most talented people and models (such as Lloyd-Ellis and Bernhardt, 2000) in which poor capital markets hurt most the least talented.

lived and endows them with the standard forward-looking preferences. In the case in which the production technology is convex and credit markets are absent, this is a special case of the model studied by Loury (1981), where he showed that there is convergence despite the credit markets being absent. Indeed, as emphasized by Casselli and Ventura (1996), the presence of effective credit markets in such an environment may actually slow or even stop convergence. The point is that poor capital markets can act as a spur to savings, because they make it more important to have one's own wealth and this effect is strongest for the poor, since capital is most productive in their hands. In contrast, when capital markets are perfect, the marginal product of capital is equalized everywhere, and the poor have no more incentive to save than the rich.[42]

When we combine this model of savings with a nonconvex monitoring technology, things change dramatically. As already noted, the rate of return on savings is now lower for the poor than for those who are somewhat richer. This is especially true of the very poor, who cannot borrow at all. As a result, the rich (or at least the middle classes, because the very rich also earn low returns on their savings) will save a higher proportion of their income than the poor, which reinforces the poverty trap.[43]

Theme 3. If savings decisions are based on future benefits, capital market imperfections can actually promote savings and convergence. However, if the technology of production or monitoring is nonconvex, the encouragement effect may operate only on the relatively wealthy. The poorest get low returns from capital and therefore will save relatively little, which may reinforce the poverty trap.

4.2. Endogenous Prices and Collective Poverty Traps

One implication of there being a poverty trap at the level of the individual is that there is also a collective poverty trap: an economy that starts with a lot of poverty will end up with a lot of poverty. *Collective poverty traps can, however, exist in models with imperfect credit markets, even when there are no individual-level poverty traps.* This point, first noted by Banerjee and Newman (1993),[44] relies on the fact that in a world where people are credit constrained, factor prices depend on the wealth distribution (because the demand for factors depends on who has how much wealth). However, the wealth distribution in any economy depends on factor prices – this two-way interaction creates the possibility of multiple steady states.

[42] As pointed out by Ghatak, Morelli, and Sjostrom (2000), it can also act as a spur to hard work, as a way to accumulate capital.

[43] This is strictly true only if we assume that the average savings rate for the economy is still β, but now the poor and the rich have different savings rates.

[44] There can also be collective poverty traps in models in which there are no credit market imperfections, if, for example, there are peer group externalities (see Durlauf, 1996).

To see exactly how this might happen, consider a variant of the model developed herein with linear production and monitoring technologies. The one factor price in that model was the cost of capital, which, so far, we took as given. To endogenize the interest rate, assume that the supply of capital comes from a fixed fraction of the population who cannot invest in the linear technology. Assume that, in every generation, a fraction μ are handicapped in this way, but this attribute is neither correlated over time nor correlated with their wealth. These people (and everyone else) do have an alternative investment possibility, which, for want of a better description, we will call "land": The return from investing in land is given by a strictly concave production function $H(\widetilde{\mu K})$, where $\widetilde{K}$ is the average amount invested in land by each investor who invests in land. Because this is a completely safe investment, it will earn the safe rate, ρ, on the marginal unit, that is, $\rho = H'(\mu\widetilde{K})$. This allows us to write $\rho = \rho(\widetilde{K})$, $\rho' < 0$. The rest of the available wealth in the economy will be invested in the linear production technology. In other words, if $\overline{W}$ is the per capita wealth in the economy, and if $\overline{K}$ denotes the average amount that each investor puts into the linear production technology, the market clearing condition for the credit market will be

$$\mu\widetilde{K} = \overline{W} - (1-\mu)\overline{K} = \overline{W} - (1-\mu)\lambda(\rho(\widetilde{K}))\overline{W}, \tag{4.1}$$

where $\lambda(\rho)$ is the optimal leverage ratio from the point of view of borrowers in an economy where the cost of capital is ρ.

How about the evolution of wealth in this economy? Given all the assumptions that have already been made, this turns out to be quite straightforward:

$$\begin{aligned}\overline{W}_{t+1} = \beta[e + H(\mu\widetilde{K}(\overline{W}_t)) + (\overline{W}_t - \mu\widetilde{K}(\overline{W}_t))\\ \times (\sigma E(p(\rho(\widetilde{K}))) - M(\rho(\widetilde{K})))],\end{aligned}$$

where $p(\rho)$ is the optimal choice of p when the cost of capital is ρ and $M(\rho)$ is the corresponding level of monitoring per unit of capital. From Equation (4.1), when $\overline{W}$ goes up, $\widetilde{K}$ also goes up, but less than proportionally because ρ goes down and λ goes up. In other words, as the economy gets richer, a higher and higher proportion of its wealth will be invested in the linear production technology. Because the net return to the linear production technology, $\sigma E(p(\rho)) - M(\rho)$, is greater than ρ, from Equation (3.7), the average return on capital may go up as a result of the shift between the two sectors. This tells us that the $\overline{W}_{t+1}(\overline{W}_t)$ map need not be concave. In particular, it can have the S shape depicted in Figure 1.4. This is most likely if it is the case that at low levels of $\overline{W}_t$ most of the capital is invested in land, but when $\overline{W}_t$ goes up beyond a certain point, the marginal product of capital invested in land falls off very quickly and, as a result, all additional capital is allocated to the alternative technology.

The S shape in Figure 1.4 generates what one might call a collective poverty trap. No individual in this economy is ever trapped in poverty because all the

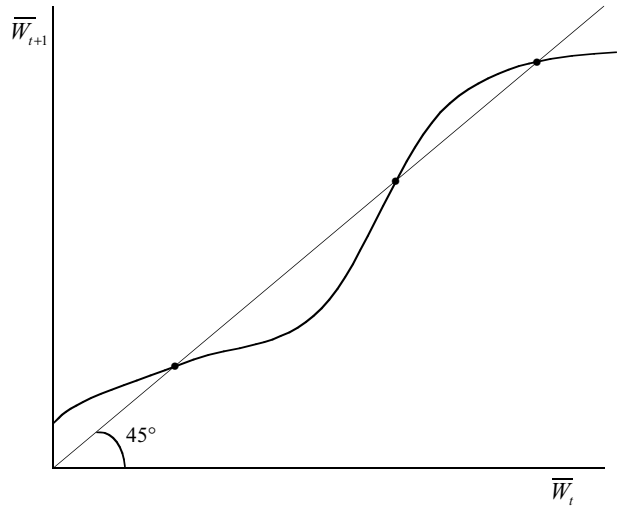


Figure 1.4.

technologies are convex. Nevertheless, economies that start poor stay poor. This is because capital is expensive in these economies and we know from our previous discussion that, when capital is expensive, it is more likely to be used wastefully (our Property 1).

This argument is an instance of a very general point: In all of these models, the distribution of wealth determines the pattern of investment, which in turn determines the demand for various factors, the factor prices, and eventually the next period's distribution of wealth. Because the effect of demand on factor prices is usually nonlinear, the map from the present distribution of wealth to the future distribution so generated will typically also be highly nonlinear, and therefore there is no reason to expect unique steady states in dynamic models with endogenous price determination.[45] Indeed, there is no presumption that these dynamic processes necessarily converge to a steady state: Aghion et al. (1999) and Aghion, Bacchetta, and Banerjee (1999) generate endogenous cycles from models of this class.

This type of argument can also be made by using other prices: Banerjee and Newman (1993) generate this type of multiplicity based on the endogeneity of wages. High-wage economies, in their argument, allow the children of the poor to become entrepreneurs, with the result that the demand for labor and wages remains high.

What is key in all of these cases is a perverse price effect: In the presence of capital market imperfections, price changes can have powerful wealth effects

[45] For a more elaborate discussion of this point, see Banerjee and Newman (1993).

that make it possible that an increase in the price actually leads to an increase in the excess demand. Thus, in the just-mentioned paper by Banerjee and Newman, an increase in the wage raises the demand for capital, whereas in the example represented in Figure 1.4, a rise in the interest rate raises the excess demand for capital.

Perverse price effects are, however, no guarantee that global convergence will fail. Aghion and Bolton (1997), who were the first to analyze a dynamic model of exactly this class (i.e., with credit market imperfections and an endogenous interest rate), had focused on the case where there was global convergence. Their point was that the process of convergence involved a Kuznets-like sequence of increasing and then decreasing inequality. Our example is inspired by the work of Piketty (1997), who showed that the Aghion–Bolton model can, under suitable parametric assumptions, generate multiple steady states.[46] The key difference seems to come from assumptions about the elasticity of demand for capital: In the Aghion–Bolton model the demand for capital is relatively elastic and therefore there are no sharp changes in the interest rate. As a result, the economy is always quite stable.[47]

Theme 4. There can be collective poverty traps even when there are no individual poverty traps. This is most likely to be the case if the supply of and demand for factors are not too price elastic.

4.3. Taking Stock: How Plausible Are Poverty Traps?

There are two parts to the answer to this question. First, are the assumptions of the poverty trap model plausible? Second, do the implications of the model correspond to something we observe? These are addressed one by one.

One key assumption of the model of the poverty trap at the level of an individual is a nonconvexity in either the production function, the monitoring function, or the bequest function. Of the three, nonconvexities in the bequest function are perhaps the easiest to document. Empirical evidence from many OECD countries supports the view that bequests are a luxury good. Only the

[46] In Piketty's example (as in our example), multiple steady states arise because high interest rates are inefficient and this inefficiency reduces the supply of capital, and therefore the interest rate goes up. Interestingly, Matsuyama (2000), who provides an alternative argument for multiple steady states in this class of models, relies on an argument that associates high interest rates with *efficient* steady states. Raising the interest rate in his model allows the poor to accumulate wealth faster (because they tend to be lenders), and this increased wealth allows more poor people to make the transition faster into being investors, raising the demand for capital and the interest rate. In other words, although both Piketty and Matsuyama build their argument on the fact that in this type of model the excess demand for capital can go up when the interest rate goes up, in Matsuyama this happens because the demand for capital goes up whereas in Piketty it is the supply that goes down.

[47] Although high interest rates are typically associated with underdevelopment – which is what Piketty's model tells us – neither of these models is intended to be "taken to the data."

richer people leave bequests of any significant size, so the bequest function, at least over a range, is clearly convex.

Nonconvexities in production are certainly very plausible. Most machines have a minimum efficient scale, and although rental markets can ameliorate this problem, they provide, at best, an imperfect substitute.[48] In the case where the good being produced is a usable education, there are several potential sources of nonconvexities: Learning the letters of the alphabet is probably useful only when it translates into the ability to read simple sentences. For this reason, the first few years of education may not generate any returns, unless instruction is continued further. This is what Card and Krueger (1992) find in U.S. data: The first five years of education have no direct return. In the United States, of course, very few people plan to get less than five years of education, but for some developing countries this may be an important nonconvexity. Angrist and Krueger (1999), once again using U.S. data, find significant jumps in the return to education at school completion and college completion. Case and Deaton (1999) also report that the relation between wages and years of schooling for black South Africans starts relatively flat and becomes steeper. However, Psacharopoulos (1994) concludes in favor of concavity, on the basis of a range of studies from all over the world, and Duflo (2001) finds an essentially linear relationship on the basis of data from Indonesia. Assessing the relative merits of these studies is beyond the scope of this paper, but it is clear that the definitive paper on the important issue of nonconvexity in education is yet to be written. Note, however, that in any case the fact that there a nonconvexity is interesting only if it is large enough to be a hurdle for a significant number of people: Ideally, we would like evidence that allows us to scale the wealth distribution among potential investors in an industry to the size of the nonconvexity in that industry. This would allow us to answer questions such as: What fraction of the population of potential investors in this industry are sufficiently poor that the nonconvexity is relevant for them?

The situation in the case of the monitoring technology is even worse because very little is actually known. Aleem reports only the average monitoring cost for the entire population, which does not tell us how the cost changes with the amount borrowed. There are certainly a priori good reasons to suspect that some part of the cost – such as the cost of meeting the first time with the potential borrower – is a fixed cost. Moreover, as already discussed, fixed costs provide a natural explanation of the observed patterns in credit markets. However, we certainly need direct evidence about the size of the costs that are incurred by the lender at different stages of his or her relationship with borrowers and how that varies across borrowers.

There is also an a priori argument against nonconvexities. Lehnert (1998) has argued that, wherever there are nonconvexities, people should participate in lotteries, which will make all of them better off and eliminate both the

[48] Machines that require careful maintenance are typically not available for rent.

nonconvexity and the poverty trap. This is not as fanciful as it might seem. Many poor people throughout the world participate in ROSCAs, which, it has been argued, are a type of private lottery designed to deal with nonconvexities.[49] However, ROSCAs typically have the feature that all participants get to "win" over a relatively short period of time (at most a year), which is feasible when the nonconvexity is of the order of magnitude of a year's savings for the average participant but not for larger nonconvexities. This may reflect the fact that people are unwilling to lose a large amount of money in a single lottery, perhaps because there is loss aversion or regret in their preferences. Moreover, the argument applies only to the nonconvexities in production and monitoring. What is described as a nonconvexity in the bequest technology is actually a simple nonlinearity and is not subject to this criticism.

There is also the possibility that as the economy becomes richer and richer (through, e.g., a process of technological upgrading), the nonconvexities will become less and less important. For this to be true, the level of the nonconvexity must grow less fast than the income of the average person: This is unlikely to be true if, for example, the nonconvex cost is a labor cost (e.g., the cost of monitoring). Moreover, the process of technological upgrading is often accompanied by an increase in the size of the fixed cost.

Nonconvexities, by themselves, do not guarantee that there is a poverty trap. What we need for an individual-level poverty trap, to put it crudely, is evidence that the map from current wealth to future wealth is steep enough to cut the 45° line more than once. This can be a tough criterion to meet. For example, Dasgupta (1993) has argued, based on evidence from studies by biologists and nutritionists of the effects of malnutrition, that the relation between the parent's health (his proxy for wealth) and the child's health tends to be highly nonlinear and typically includes a nonconvex section. Although this is important evidence, it is not enough to establish the existence of a poverty trap as interpreted here.[50] The problem is that the available estimates of the elasticity of income with respect to nutrition and health, as well as the elasticity of nutrition with respect to income, tend to be less than one,[51] suggesting that there cannot be an individual-level poverty trap.[52] In contrast, it can be argued that this is, of course, only one of many mechanisms that go into the map from current to future wealth. The question then is whether these mechanisms tend to reinforce each other or cancel each other out; note that they can cancel each other out even if there is no actual interaction between the various mechanisms,

[49] See Besley, Coate, and Loury (1993).

[50] Dasgupta (1997) suggests a more inclusive definition of a poverty trap in response to this kind of criticism.

[51] Strauss and Thomas (1993) list more than twenty studies that estimate the relation between income and calorie consumption (as a meaure of nutrition). Each one of them reports an estimate of less than 1 and most are less than 0.5. Strauss (1986) estimates the elasticity of wages with respect to nutrition in Sierra Leone and reports an elasticity of about 0.5.

[52] Srinivasan (1994) makes a similar point.

simply through a process of averaging.[53] This seems to be an important area for future research.

The other pillar of this class of theories is wealth effects on investment. The evidence on wealth effects on access to credit has already been discussed at some length, and it may be presumed that these translate into wealth effects on investment. However, it is not clear that the causal factor here is necessarily wealth rather than some correlate of wealth. More direct evidence is now available from studies of firms in the United States,[54] showing that firms that get positive cash-flow shocks invest more, even after controlling for changes in their productive opportunities. In terms of developing-country data, there is a recent study by Jeong and Townsend (2000), based on data from Thailand, that shows that the probability of owning a nonfarm business is less than 10 percent in the bottom decile and over 30 percent in the top decile.[55] It not clear, however, that the entire effect they find is attributable to the direct effect of wealth rather than the effect of characteristics that are correlated with wealth. Banerjee and Duflo (2001) use a change in directed lending policies in India as a natural experiment to estimate the effect of greater access to working capital on profits. They find that an extra *one rupee of credit increases profits, net of interest, by more than one rupee*. Duflo (2000) finds evidence of strong income effects on investment in the health of young girls in South Africa, though it is not clear that it primarily reflects the effect on access to credit. However, these studies are relatively recent, and only time can tell whether their results will turn out to be robust enough to be the foundations of a theory of development. As of now, as in many other instances, theory seems to be ahead of the evidence.

Collective poverty traps, unlike poverty traps at the level of the individual, do not directly rely on nonconvexities. Wealth effects on investment are, however, clearly crucial, as are strong price effects, which in turn rely on demand and supply for factors being relatively inelastic. Although we do have some estimates of these elasticities (though mainly from developed countries), the theory is yet to be developed in a form that would allow calibration by using these estimates.

The other approach to evidence would be to look directly for poverty traps. In other words, we could study the rates of economic mobility at different levels of wealth. Or we could look at the evolution of the wealth distribution and try to estimate from it the implied parameters of the underlying economy (including mobility rates). Robert Townsend, in joint work with a number of his coauthors, has done interesting research along these lines.[56] The problem remains, however, that we have no independent estimate of the part of mobility that arises from sources that are excluded from our model, such as genetics or

[53] For example, a series of nonconvexities at slightly different places may average to a convex map.

[54] See, for example, Fazzari, Hubbard, and Petersen (1988) and Fazzari and Petersen (1993).

[55] Evans and Jovanovic (1989) show a similar result for the United States.

[56] See Jeong and Townsend (2000).

learning.[57] For this reason, a more promising approach may be to look at the mobility patterns within a single industry, where it may be easier to calibrate some of the other sources of mobility.

There have also been attempts to use cross-country data on inequality and growth to look at this question.[58] These exercises suffer from two problems. First, a number of theories, including those discussed here, predict a relation between inequality and growth, and the data do not distinguish between these alternative channels. Second, as argued by Banerjee and Duflo (1999), there is severe danger of misspecification, because of both possible omitted variables and strong nonlinearities in the data.

4.4. Beyond Poverty Traps

Poverty traps are only the starkest form of what makes the world with imperfect credit markets interesting. As already noted, the more general phenomenon of slow convergence and limited social mobility is both interesting and important in itself. There are a number of other interesting predictions of the imperfect credit market model, some of which are now briefly sketched.

First, poor capital markets will tend to be associated with great diversity in firm size within the same industry. In the case in which there is an optimal scale of production, we would expect to see firms that are both below and above this scale – the former because they are capital starved and the latter because overall underinvestment generates rents in the industry, and as a result those who do have the capital overinvest in order to capture those rents. This appears to be consistent with the oft-remarked fact that most industries in developing countries have both large firms and a fringe of very small firms.

Second, one might expect to find some very diversified firms in economies with poor capital markets. This is because there will be rents in most industries and those who have money to invest can capture these rents. Therefore, the extent of diversification may not be guided by the usual considerations of competence and synergies. The trading companies in Japan, the Chaebols in Korea, and the managing agencies in India are potential examples of this phenomenon.

Third, industry may be clustered in certain locations that have little to recommend them as venues for that industry. This is because informal lenders prefer to lend to borrowers they can monitor easily, and this might lead them to prefer those who invest locally and in familiar industries. In Banerjee and Munshi (2001), we argue that the concentration of the knitted garment industry in Tirupur, an otherwise unknown town with poor infrastructure, is partly a result of the fact that the local population of Gounders need an outlet for their agricultural surplus.

[57] There are attempts to measure the contribution of genetics to economic mobility. See Bowles and Gintis (2001), for example.

[58] See Benabou (1996) for an excellent review of this literature.

Finally, tied transactions will be very important in this world, because tying saves on the costs of monitoring.[59] Thus, sellers will be the preferred source of credit for buyers (trade credit) and employers will be the preferred lender for workers. Cunat (2000) justifies the persistence of expensive trade credit in OECD countries (the standard rate in the United States is 44 percent) in these terms.

5. OTHER CONTRACTING PROBLEMS

Contracting problems in the land market are, not surprisingly, very similar to the problems in credit markets. The models predict that wealthier tenants will get more efficient contracts and more land.[60]

Contracting problems faced by lenders are, of course, just the other side of the contracting problems faced by borrowers. However, the models of poverty traps typically focus on the borrowers, even though it is easy to see how a very similar story would apply to lenders. In this story, the poor will not be able to lend to individual borrowers because there are fixed costs of monitoring, but intermediaries in the credit market also do not want to deal with them because there is a fixed cost of collecting money from them (e.g., they have to meet). Therefore, they earn very low returns on their savings. Rutherford (1999) documents numerous examples in which the poor accept substantial negative rates of return in order to put their savings in a safe place. Given the low return on savings, they prefer not to save and stay poor.

This story is empirically at least as plausible as any other story of the poverty trap, especially given that the poor are more likely to be lenders than borrowers. Yet there are few models of this type. Matsuyama (2000) provides one paper that does take this issue seriously, but his focus is on collective rather than individual-level poverty traps.

Contracting problems in the insurance market, by contrast, can be quite unlike those in credit markets. Insurance is a key market for the poor because they may be extremely vulnerable to even small changes in their consumption (i.e., they are likely to be more risk averse than the rich).[61] This is consistent with the large and growing literature that shows that the poor enter into many sophisticated arrangements in order to limit their risk exposure.[62] As noted by Newman (1995), this is not inconsistent with the possibility that the poor are treated better than the rich in the optimal insurance contract.[63] The reason comes from precisely the fact that the poor have more to lose. This means the threat of

[59] This argument goes back to Bardhan (1983).

[60] This phenomenon is well known in the empirical literature on tenancy and goes under the name of "tenancy ladder."

[61] See, e.g., Deaton (1989).

[62] See, e.g., Udry (1994) and Townsend (1995).

[63] In the sense that the rich would want to have the insurance contract that the poor get in equilibrium.

even a small loss can give good incentives to the poor, making it easier to give them good insurance as well. Newman goes on to show that under reasonable assumptions, this effect can be so strong that the poor would be prepared to take on risky and profitable projects that the rich will avoid. This is, of course, quite different from the predictions of the credit market model – there, it is the rich who do the profitable projects. It is clearly a force toward convergence, and it explains why there is no poverty trap in the model of Banerjee and Newman (1991) on risk bearing. Of course, this kind of result depends crucially on the setup. We would get the opposite result if the insurance market was entirely absent, perhaps because the fixed costs of enforcing such a contract are too high.[64] Then there could easily be a poverty trap; in such a model the poor would underinvest because investment is risky and they are unwilling to bear any risk.[65]

The contrast between the credit market case and the insurance case does not end here. Keeping the insurance contract fixed, increasing risk exposure (e.g., by weakening social protection) and increasing risk aversion hurt the poor in the insurance context but may actually help them get more credit (because it makes it easier for the lender to "punish" them for defaulting).

The source of this conflict lies in the basic premises of these two narratives: In the insurance market view, the emphasis is on the vulnerability of the poor, that is, on the fact that they cannot afford any losses; in the credit market view, the poor are seen as unreliable because they have too little to lose and therefore cannot be punished for defaulting. These views are not necessarily inconsistent, for example, because it may be very costly for the lender to inflict losses on the poor, even though losses hurt the poor a lot.[66] Or, the fact that the poor have too little to lose, and therefore are unable to invest, might make the slightly less poor extremely averse to the risk of becoming poor and therefore *unwilling to invest*.[67] In other words, it is possible that the credit market and insurance market views reinforce each other, but it remains that the tension between them is real and far from being resolved.

Contracting in product markets is less studied in the context of less developed countries (LDCs). However, as the share of quality-sensitive products in world demand grows (as it has in recent years because of the growth in the new economy), sellers in LDCs will have to be able to assure buyers that they are getting the desired quality. Because quality is not easy to contract on ex ante, this will raise the importance of appropriate contract design in product markets for LDCs. It will also make reputation and brand names much more valuable, making it harder for new entrants. One case where

[64] See Kanbur (1979) and Kihlstrom and Laffont (1979) for results of this type.

[65] For a simple model of a poverty trap of this type, see Banerjee (2000). Morduch (1995), and Walker and Ryan (1990) provide some suggestive evidence for the view that the poor are discouraged by risk from taking up the most profitable opportunities.

[66] See Banerjee (2000) for a more elaborate discussion of these issues.

[67] As in Banerjee (2000).

this has already happened is the Indian software industry, where Banerjee and Duflo (2000) show that the more reputed firms get both better contracts and more rewarding projects.

6. THE VIEW OF POLICY

If the evidence and arguments listed herein do one thing, it is to challenge the hegemony of the complete markets Arrow–Debreu model as the basis for policy thinking. The usual view in economics seems to be that the complete markets model provides the natural framework, modified perhaps by acknowledging some limited role for transactions costs. My view is that there are many important contexts in which these transaction costs are so large, and the consequent deviation from the complete markets model so glaring, that it is better to abandon the complete markets model, except inasmuch as it provides a useful intellectual point of reference. To take an example, in the market studied by Aleem, the transaction cost of about fifty cents on a dollar clearly swamps the interest rate paid to the ultimate lender (the depositor in a bank, who gets ten cents). However, this is only the observable part of the transaction cost. Then there is the cost of missed opportunities, because in addition to these high rates the lenders probably impose credit limits and, moreover, some people are completely excluded from the market because no one knows or trusts them.[68] Then there are dynamic costs: The fact that the current borrower underinvests (or earns low net returns on his or her investment) means that the borrower's son or daughter will also be poor and will also not be able to take advantage of his or her opportunities and talents. Finally, there are general equilibrium effects: Inefficiency in investment means low wages and slow capital accumulation today, both of which contribute to poverty and inefficiency tomorrow.

The rejection of the complete markets model should not be seen as a justification for old-fashioned dirigiste policies. There are, of course, good reasons to worry about the deliberate misuse of these policies. But perhaps of equal importance is that the recognition that markets often fail does not automatically imply that we should pursue antimarket policies. To take an example, trade protection may be particularly bad if capital markets are imperfect, because it reinforces the capital market frictions that slow down the flow of capital toward its best possible uses.

What does emerge from the analysis herein is the need to build policies that recognize the market failures that are most important in the particular context. Thus, assessing the growth impact of trade policies without taking account of the distributional impact of these policies is self-defeating, because the distributional impact will frame the future pattern of investment.

[68] There is obviously a trade-off here. Some lenders may limit themselves to borrowers they know very well, which brings down the direct transaction cost but increases the costs coming from exclusion and missed opportunities.

The imperative of taking specific market failures seriously when making policy is obviously rather vague. To give it some more content, I now discuss some simple examples of how thinking about the world in this way feeds into specific policy recommendations.

The first example, already alluded to, comes from credit and insurance markets. As we saw, the implications of better social protection tend to be very different in models of credit and insurance. This must be taken into account in designing social protection mechanisms.

Second, it is clearly important to try to reduce the cost of credit to the poor. One idea that has received a lot of currency is to make use of peer monitoring and screening by peers through micro finance institutions.[69] This has two related advantages. First, members of one's peer group may be better at monitoring and screening; this reduces the cost of monitoring. Second, the usual arrangements involve mutual monitoring on a quid pro quo basis. Therefore, the interest rate does not have to be raised in order to pay for the cost of monitoring. Lower interest rates, as we already saw, generate better incentives and therefore less monitoring is needed.[70] Other possible interventions include trying to develop systems of credit rating and centralizing credit histories so that the credit markets become less segmented and borrowers have access to the cheapest sources of credit. Developing better systems for the recording of property ownership (so that the assets that the poor have can be used as collateral), and a court system that resolves property disputes quickly and effectively (so that lenders believe that they can collect on the collateral), will also help the poor. Helping the poor to develop credit histories and helping them to learn to deal with the financial system more generally (e.g., by keeping proper accounts) is yet another potentially fruitful avenue.[71]

Third, giving a central role to issues of credit access gives a new urgency to the old policy prescription of encouraging savings. It is easy to see that one way to get out of poverty traps is to raise the savings rate. In Figures 1.1, 1.2, 1.3, and 1.4, this pushes up the map from current wealth to future wealth and thereby makes poverty traps (both individual and collective) less likely. More generally, as already discussed, the poor may save too little (relative to the social optimum) when capital markets are imperfect, both because the rates paid to depositors tend to be too low (because lending is difficult) and because they do not have the wherewithal to become investors and therefore put their savings to the best possible use. This is compounded by the fact that intermediaries in the credit market may bypass them for the simple reason that given the small volume

[69] See Banerjee, Besley, and Guinnane (1994) for a model of lending based on peer monitoring and Ghatak (2000) for a model of screening by peers.

[70] This argument relies on the assumption that the time spent on monitoring has no alternative cash-generating use.

[71] This is one way to interpret the contribution of many of the nongovernmental organizations that work in micro credit but focus on lending to individuals rather than groups. BRI in Indonesia is a well-known example.

of their savings, the fixed cost of collecting savings may swamp any potential returns from investing them. Subsidizing access to savings opportunities may therefore be a powerful weapon for helping the poor.

Fourth, thinking of the underlying contract theory clearly gives us a very different perspective on land reform and tenancy reform. In particular, it tells us that it may be possible to achieve many of the desirable productivity effects of such reforms without actually changing any land rights. One example of such an intervention would be a program that improves the outside options of the tenants, such as an employment guarantee scheme.[72] As the tenant's outside option improves, he or she will be rewarded more, and that can lead to an improvement in his or her incentives.[73]

Finally, if our prediction about the increasing importance of product market contracting in developing countries is borne out, some of these countries will have to make significant policy changes so that they are not left out. In part, this will involve strengthening the court system, but in part it will also require other innovations such as helping domestic companies build a reputation or making it possible for domestic producers to enter into strategic partnerships with reputable multinational companies in order to benefit from their reputation.[74]

7. CONCLUSION

I have made the choice in this survey of focusing on one rather specific topic in order to explain better the logic of how research in this area of development economics has evolved over the past twenty years, and to pinpoint key gaps in our knowledge in this relatively well-studied area. To conclude, let me now say something about the broader agenda.

The one most important limitation of my survey is its focus on the level, distribution, and growth of output as the main outcome of interest. Development economics is much broader: There is a large and important literature that attempts to explain the existence and persistence of institutions such as sharecropping, the village moneylender, ROSCAs, community-based lending networks, cooperatives, and so on, based on contract theoretic arguments.[75] There is also a long tradition that argues that institutions act as an independent force in the

[72] The strategy of improving outside options is called an empowerment strategy in Banerjee, Gertler, and Ghatak (2000), who also provide a formal analysis of how these strategies work.

[73] For a more detailed analysis of land reforms and alternatives to land reform, see Banerjee (1999).

[74] In many cases this may not involve much more than allowing foreign equity participation in domestic companies. However, in cases where countries suffer from a collective reputation problem, there may be a case for a more substantial intervention: penalizing domestic companies that fail to meet some quality standard may be one way of achieving this (see Tirole, 1996, for a model of collective reputation).

[75] Bardhan (1989) and Greif (1997) are two important sources for this literature. Legros and Newman (1996), Wells (1999), and Prescott and Townsend (2000) are interesting recent examples of this style of work.

economy and influence economic outcomes. The Lewis model, which argued that the particular structure of the implicit contract in family farms restrains migration and growth, is perhaps the most well-known example of a model of this type. Although there was a period when economists seemed to take the extreme Coasian view that inefficient institutions should not exist, there is now a clearer understanding that there is no good reason why institutions should be set up with an eye toward global optimality. Banerjee and Newman (1998) present an example where a locally efficient institution leads to inefficient global outcomes.[76]

What are still rare are dynamic models where the institutions themselves evolve. Banerjee and Newman (1993) present one such model, aimed at explaining the evolution of the economic institutions of capitalism (large capitalist firms as against self-employment) as a result of change in the wealth distribution.[77] Greenwood and Jovanovic (1990) present a model where the financial sector evolves with growth.

Even rarer are models where institutions not only evolve but actually have feedback effects on the rest of the economy. Some recent examples include Acemoglu and Robinson (1998), which looks at the evolution of the franchise, Acemoglu and Zilibotti (1997), which focuses on the financial sector, and Banerjee and Newman (1998), which looks at the evolution of the modern sector. However, there are not many others, even though it seems clear that this is the process that development economists would like to capture.

Much research has been done in the past twenty years. If it has achieved anything, it is to make us aware of where we would like to be and what has to be done. Now, I hope, it is only a matter of time.

ACKNOWLEDGMENTS

This paper grew out of a lecture I gave at the World Congress of the Econometric Society in August, 2000. I am grateful to Philippe Aghion and Esther Duflo for their comments and to Marko Tervio for his help with simulations.

References

Acemoglu, D. and J. Robinson (1998), "Why Did the West Extend the Franchise? Democracy, Inequality, and Growth in Historical Perspective," Discussion Paper 1797, Centre for Economic Policy Research.

Acemoglu, D. and F. Zilibotti (1997), "Was Prometheus Unbound by Chance? Risk, Diversification, and Growth," *Journal of Political Economy*, 105(4), 709–751.

[76] Douglas North has been the most influential exponent of this general position in recent years. See Grief (1994) for an interesting and nuanced statement of this view.

[77] See also the related work by Lloyd-Ellis and Bernhardt (2000).

Aghion, P., P. Bacchetta, and A. Banerjee (1999), "Capital Markets and the Instability of Open Economies," unpublished, University College of London and Harvard University.

Aghion, P., A. Banerjee, and T. Piketty (1999), "Dualism and Macroeconomic Volatility," *The Quarterly Journal of Economics*, 114(4), 1359–1397.

Aghion, P. and P. Bolton (1997), "A Theory of Trickle-Down Growth and Development," *The Review of Economic Studies*, 64(2), 151–172.

Aleem, I. (1990), "Imperfect Information, Screening, and the Costs of Informal Lending: A Study of a Rural Credit Market in Pakistan," *World Bank Economic Review*, 3, 329–349.

Angrist, J. and A. Krueger (1999), "Empirical Strategies in Labor Economics," in *Handbook of Labor Economics*, Vol. 3, (ed. by O. Ashenfelter and D. Card), New York: Elsevier Scientific, 1277–1366.

Banerjee, A. (1992), "A Simple Model of Herd Behavior," *Quarterly Journal of Economics*, 107(3), 797–817.

Banerjee, A. (1999), "Prospects and Strategies for Land Reforms," mimeo, MIT.

Banerjee, A. (2000), "The Two Poverties," *Nordic Journal of Political Economy*, 26(2), 129–141.

Banerjee, A. (2001), "What Do We Know about Credit Markets?" manuscript in preparation, MIT.

Banerjee, A., T. Besley, and T. Guinnane (1994), "Thy Neighbor's Keeper: The Design of a Credit Cooperative with Theory and a Test," *The Quarterly Journal of Economics*, 109(2), 491–515.

Banerjee, A. and E. Duflo (1999), "Inequality and Growth: What Can the Data Say?" mimeo, MIT.

Banerjee, A. and E. Duflo (2000), "Reputation Effects and the Limits of Contracting," *Quarterly Journal of Economics*, 115(3), 989–1017.

Banerjee, A. and E. Duflo (2001), "The Nature of Credit Constraints: Evidence from Indian Bank," mimeo, MIT.

Banerjee, A., P. Gertler, and M. Ghatak (2000), "Empowerment and Efficiency: Tenancy Reform in West Bengal," *Journal of Political Economy*, 110(2), 239–280.

Banerjee, A. and K. Munshi (2001), "How Efficiently Is Capital Allocated? "Evidence from the Knitted Garment Industry in Tirupur," mimeo, MIT.

Banerjee, A. and A. Newman (1991), "Risk-Bearing and the Theory of Income Distribution," *The Review of Economic Studies*, 58(2), 211–235.

Banerjee, A. and A. Newman (1993), "Occupational Choice and the Process of Development," *Journal of Political Economy*, 101(2), 274–298.

Banerjee, A. and A. Newman (1994), " Poverty, Incentives, and Development," *American Economic Review Papers and Proceedings*, May, 64(2), 211–215.

Banerjee, A. and A. Newman (1998), "Information, the Dual Economy, and Development," *The Review of Economic Studies*, 65(4), 631–653.

Bardhan, P. (1983), "Labor-Tying in a Poor Agrarian Economy: A Theoretical and Empirical Analysis," *Quarterly Journal of Economics*, 98(3), 501–514.

Bardhan, P. (Ed.) (1989), *The Economic Theory of Agrarian Institutions*. Oxford: Oxford University Press.

Benabou, R. (1996), "Inequality and Growth," in *NBER Macroeconomics Annual 1996*, Cambridge, MA: MIT Press, 11–74.

Besley, T. and A. Case (1994), "Diffusion as a Learning Process: Evidence from HYV Cotton," Research Program in Development Studies Discussion Paper 174, Princeton University.

Besley, T., S. Coate, and G. Loury (1993), "The Economics of Rotating Savings and Credit Associations," *American Economic Review*, 83(4), 792–810.

Bhaduri, A. (1977), "On the Formation of Usurious Interest Rates in Backward Agriculture," *Cambridge Journal of Economics*, 1(4), 341–352.

Bhagwati, J. (1982), "Directly Unproductive, Profit-Seeking (DUP) Activities," *Journal of Political Economy*, 90(5), 988–1002.

Bottomley, A. (1963), "The Cost of Administering Private Loans in Underdeveloped Rural Areas," *Oxford Economic Papers*, 15(2), 154–163.

Bowles, S. and H. Gintis (2001), "The Intergenerational Transmission of Economic Status: Education, Class, and Genetics," in *Genetics, Behavior, and Society*, a volume in Neil Smelser and Paul Baltes, eds., *International Encyclopedia of the Social and Behavioral Sciences* (ed. by M. Feldman), Oxford: Elsevier.

Caballero, R. and M. Hammour (2000), "Creative Destruction in Development: Institutions, Crises, and Restructuring," mimeo, MIT.

Card, D. and A. Krueger (1992), "Does School Quality Matter? Returns to Education and the Characteristics of Public Schools in the United States," *Journal of Political Economy*, 100(1), 1–40.

Case, A. and A. Deaton (1999), "School Inputs and Educational Outcomes in South Africa," *Quarterly Journal of Economics*, 114(3), 1047–1084.

Casselli, F. and J. Ventura (1996), "A Representative-Agent Theory of Income Distribution," mimeo, MIT.

Cheung, S. N. S. (1968), "Private Property Rights and Sharecropping," *Journal of Political Economy*, 76(6), 1107–1122.

Cole, H., G. Mailath, and A. Postlewaite (1992), "Social Norms, Savings Behavior, and Growth," *Journal of Political Economy*, 100(6), 1092–1125.

Cunat, V. (2000), "Trade Credit: Suppliers as Debt Collectors and Insurance Providers," Discussion Paper, LSE-FMG.

Dasgupta, A. (1989), *Reports on Informal Credit Markets in India: Summary*. New Delhi: National Institute of Public Finance and Policy.

Dasgupta, P. (1993), *An Inquiry into Well-Being and Destitution*, Oxford: Oxford University Press.

Dasgupta, P. (1997), "Poverty Traps," in *Advances in Economics and Econometrics: Theory and Applications–Seventh World Congress*, Vol. 2, (ed. by D. Kreps and K. Wallis), Econometric Society Monographs, No. 27. Cambridge: Cambridge University Press, 114–159.

Dasgupta, P. and D. Ray (1986), "Inequality as a Determinant of Malnutrition and Unemployment: Theory," *Economic Journal*, 96, 1011–1034.

Deaton, A. (1989), "Saving in Developing Countries: Theory and Review," in *Proceedings of the World Bank Annual Conference on Development Economics 1989* (ed. by S. Fischer and D. de Tray), Washington, DC: World Bank, 61–96.

Diamond, D. (1989), "Reputation Acquisition in Debt Markets," *Journal of Political Economy*, 97(4), 828–862.

Duflo, E. (2000), "Grandmothers and Granddaughters: Old Age Pension and Intra-Household Allocation in South Africa," NBER Working Paper 8061.

Duflo, E. (2001), "Schooling and Labor Market Consequences of School Construction in Indonesia: Evidence from an Unusual Policy Experiment," *American Economic Review*, 91(4), 795–813.

Durlauf, S. (1996), "A Theory of Persistent Income Inequality," *Journal of Economic Growth*, 1(1), 75–93.

Evans, D. and B. Jovanovic (1989), "An Estimated Model of Entrepreneurial Choice under Liquidity Constraints," *Journal of Political Economy*, 97(4), 808–827.

Fafchamps, M. (2000), "Ethnicity and Credit in African Manufacturing," *Journal of Development Economics*, 61, 205–235.

Fazzari, S. and B. Petersen (1993), "Working Capital and Fixed Investment: New Evidence on Financial Constraints," *Rand Journal of Economics*, 24(3), 328–342.

Fazzari, S., G. Hubbard, and B. Petersen (1988), "Financing Constraints and Corporate Investment," *Brookings Papers on Economic Activity*, (1), 141–195.

Galor, O. and J. Zeira (1993), "Income Distribution and Macroeconomics," *Review of Economic Studies*, 60(1), 35–52.

Ghatak, M. (2000), "Screening by the Company You Keep: Joint Liability Lending and the Peer Selection Effect," *Economic Journal*, 110(465), 601–631.

Ghatak, M., M. Morelli, and T. Sjostrom (2000), "Dynamic Incentives, Occupational Mobility, and the American Dream," mimeo, University of Chicago.

Ghatak, S. (1976), *Rural Money Markets in India*. New Delhi, India: MacMillan Company of India.

Ghate, P. (1992), *Informal Finance: Some Findings from Asia*. Oxford: Oxford University Press for the Asian Development Bank.

Gill, A. and U. C. Singh (1997), "Financial Sector Reforms, Rate of Interest, and the Rural Credit Markets: The Role of Informal Lenders in Punjab," *Indian Journal of Applied Economics*, 6(4), 37–65.

Greenwood, J. and B. Jovanovic (1990), "Financial Development, Growth, and the Distribution of Income," *Journal of Political Economy*, 98(5), 1076–1107.

Greif, A. (1994), "Cultural Beliefs and the Organization of Society: A Historical and Theoretical Reflection on Collectivist and Individualist Societies," *Journal of Political Economy*, 102(5), 912–950.

Greif, A. (1997), "Microtheory and Recent Developments in the Study of Economic Institutions through Economic History," in *Advances in Economics and Econometrics: Theory and Applications–Seventh World Congress* Vol. 2., (ed. by D. Kreps and K. Wallis), Econometric Society Monographs, No. 27. Cambridge: Cambridge University Press, 79–113.

Hart, O. and J. Moore (1994), "A Theory of Debt Based on the Inalienability of Human Capital," *Quarterly Journal of Economics*, 109(4), 841–879.

Holmstrom, B. and J. Tirole (1996), "Modeling Aggregate Liquidity," *American Economic Review*, 86(2), 187–191.

Irfan, M., G. M. Arif, S. M. Ali, and H. Nazli (1999), "The Structure of Informal Credit Market in Pakistan," Research Report 169, Islamabad: Pakistan Institute of Development Economics.

Jaffee, D. and T. Russell (1976), "Imperfect Information, Uncertainty, and Credit Rationing," *Quarterly Journal of Economics*, 90(4), 651–666.

Jeong, H. and R. Townsend (2000), "An Evaluation of Models of Growth and Inequality," mimeo, University of Chicago.

Johnson, D. G. (1950), "Resource Allocation under Share Contracts, " *Journal of Political Economy*, 58(2), 111–123.

Kanbur, S. (1979), "Of Risk Taking and the Personal Distribution of Income," *Journal of Political Economy*, 87, 769–797.

Kihlstrom, R. and J. Laffont (1979), "A General Equilibrium Entrepreneurial Theory of Firm Formation Based on Risk Aversion," *Journal of Political Economy*, 87, 719–748.

Kiyotaki, N. and J. Moore (1997), "Credit Cycles," *Journal of Political Economy*, 105(2), 211–248.

Kremer, M. (1993), "The O-Ring Theory of Economic Development," *Quarterly Journal of Economics*, 108(3), 551–575.

Krueger, A. (1974), "The Political Economy of the Rent-Seeking Society," *American Economic Review*, 64(3), 291–303.

Legros, P. and A. Newman (1996), "Wealth Effects, Distribution, and the Theory of Organization," *Journal of Economic Theory*, 70(2), 312–341.

Lehnert, A. (1998), "Asset Pooling, Credit Rationing, and Growth," Finance and Economic Discussion Paper 1998-52, Board of Governors of the Federal Reserve System.

Lehnert, A., E. Ligon, and R. Townsend (1999), "Liquidity Constraints and Incentive Contracts," *Macroeconomic Dynamics,* 3(1), 1–47.

Lewis, W. A. (1954), "Economic Development with Unlimited Supplies of Labor," *The Manchester School*, 22, 139–191.

Lloyd-Ellis, H. and D. Bernhardt (2000), "Enterprise, Inequality, and Economic Development," *The Review of Economic Studies*, 67(1), 147–168.

Loury, G. (1981), "Intergenerational Transfers and the Distribution of Earnings," *Econometrica*, 49(4), 843–867.

Matsuyama, K. (2000), "Endogenous Inequality," *Review of Economic Studies*, 67(4), 743–759.

McMillan, J. and C. Woodruff (1999), "Interfirm Relationships and Informal Credit in Vietnam," *Quarterly Journal of Economics*, 114(4), 1285–1320.

Moav, O. (1999), "Income Distribution and Macroeconomics: Convex Technology and the Role of Intergenerational Transfers," mimeo, Hebrew University.

Mookherjee, D. and D. Ray (2000), "Persistent Inequality," mimeo, Boston University.

Morduch, J. (1995), "Income Smoothing and Consumption Smoothing," *Journal of Economic Perspectives*, 9(3), 103–114.

Munshi, K. (2000), "Social Learning in a Heterogeneous Population: Technology Diffusion in the Indian Green Revolution," mimeo, University of Pennsylvania.

Murphy, K., A. Shleifer, and R. Vishny (1989), "Industrialization and the Big Push," *Journal of Political Economy*, 97(5), 1003–1026.

Murshid, K. (1992), "Informal Credit Markets in Bangladesh Agriculture: Bane or Boon?" in *Sustainable Agricultural Development: The Role of International Cooperation: Proceedings of the 21st International Conference of Agricultural Economists*, Dartmouth: Ashgate.

Newman, A. (1995), "Risk-Bearing and 'Knightian' Entrepreneurship," mimeo, Columbia University.

Nurske, R. (1953), *Problems of Capital Formation in Underdeveloped Countries*, New York: Oxford University Press.

Paulson, A. and R. Townsend (2000), "Entrepreneurship and Liquidity Constraints in Rural Thailand," mimeo, Northwestern University.

Piketty, T. (1997), "The Dynamics of the Wealth Distribution and the Interest Rate with Credit Rationing," *The Review of Economic Studies*, 64(2), 173–189.

Psacharopoulos, G. (1994), "Returns to Investment in Education: A Global Update," *World Development*, 22(9), 1325–1343.

Prescott, E. and R. Townsend (2000), "Inequality, Risk Sharing, and the Boundaries of Collective Organizations," mimeo, Federal Reserve Bank of Richmond.

Rosenstein-Rodan, P. (1943), "Problems of Industrialization of Eastern and South-Eastern Europe," *Economic Journal*, 53, 202–211.

Rutherford, S. (1999), "The Poor and Their Money: an Essay about Financial Services for Poor People," unpublished manuscript.

Schultz, T. (1964), *Transforming Traditional Agriculture*. New Haven: Yale University Press.

Srinivasan, T. N. (1994), "Destitution: A Discourse," *Journal of Economic Literature*, 32(4), 1842–1855.

Stiglitz, J. (1974), "Incentives and Risk Sharing in Sharecropping," *The Review of Economic Studies*, 41(2), 219–255.

Stiglitz, J. (1990), "Peer Monitoring and Credit Markets," *World Bank Economic Review*, 4(3), 351–366.

Stiglitz, J. and A. Weiss (1981), "Credit Rationing in Markets with Imperfect Information," *American Economic Review*, 71(3), 393–410.

Strauss, J. (1986), "Does Better Nutrition Raise Farm Productivity?" *Journal of Political Economy*, 94(2), 297–320.

Strauss, J. and D. Thomas (1993), "Human Resources: Empirical Modeling of Household and Family Decisions," in *Handbook of Development Economics*, (ed. by T. N. Srinivasan and J. Behrman), Amstersdam: North-Holland Elsevier.

Swaminathan, M. (1991), "Segmentation, Collateral Undervaluation, and the Rate of Interest in Agrarian Credit Markets: Some Evidence from Two Villages in South India," *Cambridge Journal of Economics*, 15, 161–178.

Timberg, T. and C. V. Aiyar (1984), "Informal Credit Markets in India," *Economic Development and Cultural Change*, 33(1), 43–59.

Tirole, J. (1996), "A Theory of Collective Reputations (with Applications to the Persistence of Corruption and to Firm Quality)," *Review of Economic Studies*, 63(1), 1–22.

Townsend, R. (1995), "Financial Systems in Northern Thai Villages," *The Quarterly Journal of Economics*, 110(4), 1011–1046.

Udry, C. (1994). "Risk and Insurance in a Rural Credit Market: An Empirical Investigation in Northern Nigeria," *Review of Economic Studies*, 61(3), 495–526.

Walker, T. and J. Ryan (1990), *Village and Household Economies in India's Semi-Arid Tropics*, Baltimore: Johns Hopkins University Press.

Wells, R. (1999), "Information, Authority and Internal Governance of the Firm," mimeo, Princeton University.

CHAPTER 2

Factor Models in Large Cross Sections of Time Series

Lucrezia Reichlin

1. MOTIVATION

Business cycles are characterized by two features: comovements and regular phases of expansion and depression. Comovements are observed between aggregate variables – output and inflation, for example – and between disaggregates – individual consumption and regional output, for example. The time-series literature has typically analyzed these two characteristics in a separate way. Starting with the seminal contribution of Burns and Mitchell (1946), a huge amount has been written on the "regularity" of cycles, asymmetries, and nonlinearities, on the basis of estimation of aggregate output or few relevant macroeconomic time series. A separate literature has addressed the issue of comovements, typically between few key aggregate time series and typically concentrating on long-run comovements (cointegration). Behind this literature, there is the implicit idea that the essential characteristics of the business cycle are captured by few relevant aggregate variables and that the information contained in disaggregate time series or in all the potentially available aggregate time series is not particularly useful to understand macroeconomic behavior. This is also the implicit idea behind vector autoregression (VAR) modeling, where the propagation of "identified" aggregate shocks is analyzed in models typically containing a small number of variables.

In contrast, there is a large number of econometric studies that analyze the behavior of many consumers or many firms in order to understand the microeconomic mechanisms behind fluctuations. In these studies the cross section is typically large and the time-series dimension either absent or small. Economic theory is sufficiently heterogeneous so as not to give us clear guidance on what is the level of aggregation relevant for macroeconomic questions and on what is the appropriate stochastic dimension for macroeconomic models. In the past fifteen years, the taste of the profession has been oriented toward the dynamic analysis of "small" models, but central banks and statistical institutes are still using macroeconometric models containing a large number of time series. Modern macroeconomic theory is based on the representative agent assumption, but macroeconomic empirics is mostly based on aggregate data. What is the cost of

simplicity? Are we losing valuable information by working with econometric models containing few aggregate variables? How detailed do our models have to be to have a chance to provide the essential information on the macroeconomy?

To try to answer these questions, we must develop econometric models that (a) are able to handle the analysis of many time series by reducing the number of the essential parameters to estimate; (b) can provide an answer on what is the relevant stochastic dimension of a large economy, that is, on how many aggregate shocks are needed to study the macroeconomy that emerges from the behavior of many agents; and (c) can help us to identify these (possibly few) shocks and study the propagation mechanism through agents or through geographical space. This is what will help to bridge the gap between purely time-series studies and the cross-sectional approach.

A natural starting point is the dynamic index (factor) model of Sargent and Sims (1977), Geweke (1977), and Geweke and Singleton (1981). In this framework, the dynamic of individual variables is represented as the sum of a component that is common to all variables in the economy and an orthogonal idiosyncratic residual. This approach is briefly reviewed in Section 3. Where our survey really starts, however, is Section 4, where the recent literature that has developed the index approach is analyzed.

The new developments are in different directions. First, the model is adapted to the analysis of large cross sections (n large). This is partly a purely econometric development because it deals with the issue of finding consistent estimators to the common components of each variable in the panel as n and T become larger and investigates the required relative rates of n and T at which consistency is achieved. Partly, it is a development in identification and representation theory that is deeply related with fundamental macroeconomic questions. If there are comovements in the economy, few macroeconomic shocks should explain most of the variance of relevant variables, and, if comovements characterize cyclical fluctuations, they should be observed at business cycle frequencies (corresponding to cycles of periods between three and ten years, say). Comovements imply fewer shocks than variables and therefore – loosely speaking – a factor model type of structure. For a given n, this implies conditions on the spectral density of the observations. Now, imagine that more and more information becomes available (larger n, i.e., more variables and/or more disaggregate information). If the few shocks are "pervasive," that is, they remain common to all variables in a progressively larger panel, this implies conditions on the spectral density of the observations as n becomes large, which can be exploited to define precisely the notion of common and idiosyncratic sources of dynamics. Finally, if the data support a structure in which many n variables are led by few q macro shocks, the information contained in the n variables should help to identify the q shocks, providing insights for the use of generalized factor models for policy analysis.

2. PLAN OF THE PAPER

This paper is not a survey on factor analysis. Rather, it is a survey of the relevant papers in the small but growing recent literature that analyzed dynamic factor

models in large cross sections. I discuss the relation between the latter and the static factor approach, which was developed for the study of financial data. The problems that are covered are the identification and representation of the model in population, estimation, and consistency rates, forecasting, identification of aggregate shocks, and individual propagation mechanisms. Section 3 establishes the background by briefly reviewing classical dynamic factor analysis. Section 4 reviews the main results for factor models in large cross sections. Section 5 outlines open research questions, and Section 6 reports results from some empirical applications.

3. INDEX MODELS FOR A GIVEN CROSS SECTION

The dynamic factor (index) model was introduced in macroeconomics by Sargent and Sims (1977) and Geweke (1977). Other relevant papers in the early literature are those by Geweke and Singleton (1981) and Watson and Engle (1983).

The $n \times 1$ vector of stationary variables is represented as the sum of two orthogonal components:

$$\begin{aligned} \mathbf{X}_t &= \sum_{k=-\infty}^{\infty} \mathbf{A}_k \mathbf{f}_{t-k} + \boldsymbol{\xi}_t \\ &= \mathbf{A}(L)\mathbf{f}_t + \boldsymbol{\xi}_t, \end{aligned} \tag{3.1}$$

where the $\mathbf{A}_k$ are $n \times q$ matrices, $\mathbf{f}_t$ is a stationary stochastic process of dimension $q \times 1$ and diagonal variance–covariance matrix, and $\boldsymbol{\xi}_t$ is stationary of dimension $n \times 1$ with diagonal spectral density matrix (its elements are orthogonal at all leads and lags). Both processes are allowed to be temporally correlated.

We can rewrite the model so as to express the common component $\boldsymbol{\chi}_t = \mathbf{A}(L)\mathbf{f}_t$ in terms of uncorrelated common shocks. Consider a moving average representation of $\mathbf{f}_t$, $\mathbf{f}_t = \mathbf{D}(L)\mathbf{u}_t$; we can rewrite (3.1) as

$$\mathbf{X}_t = \mathbf{B}(L)\mathbf{u}_t + \boldsymbol{\xi}_t, \tag{3.2}$$

where the common component, $\boldsymbol{\chi}_t = \mathbf{A}(L)\mathbf{D}(L)\mathbf{u}_t = \mathbf{B}(L)\mathbf{u}_t$, is expressed in terms of a q-dimension white noise (which we normalize so as to have unit variance).

The model implies that covariation among the observable variables Xs is due entirely to the common effect of the q latent factors, whereas variation of an individual variable X_{it} is due to the variation of the specific variable ξ_{it} as well as the variation and covariation of the common factor.

The restrictions on the covariance properties of the data are best understood by writing the spectral density matrix of the Xs (see Appendix A for an introduction to the spectral representation of stationary time series and the spectral density). Under the assumptions just given, the spectral density of $\mathbf{X}_t$ can be written as

$$\boldsymbol{\Sigma}_x(\theta) = \mathbf{B}(e^{-i\theta})\tilde{\mathbf{B}}(e^{-i\theta}) + \boldsymbol{\Sigma}_\xi(\theta), \quad |\theta| \leq \pi, \tag{3.3}$$

where $\mathbf{B}(e^{-i\theta})$ denotes the Fourier transform of the function $\mathbf{B}(L)$ and $\tilde{\mathbf{B}}(e^{-i\theta})$ its conjugate transpose (from now on the tilde will indicate conjugation and transposition). The model implies that the spectral density of the common shocks, $\mathbf{\Sigma}_u(\theta)$, is equal to the $q \times q$ identity matrix, $\mathbf{I}_q$, and that the spectral density of $\boldsymbol{\xi}_t$, $\mathbf{\Sigma}_\xi(\theta)$, is diagonal.

The literature has analyzed three problems: identification of the component, identification of the coefficients, and estimation.

1. Identification of the component: Because the components are not observable, prior to estimation, identification conditions have to be spelled out. In the general case in which $n \gg q$, model (3.2) is overidentified, the components $\boldsymbol{\chi}_t$ and $\boldsymbol{\xi}_t$ can be estimated by maximum likelihood techniques, and overidentified restrictions can be tested by standard likelihood ratio tests.
2. Estimation of the component: Estimation is usually performed under a normality hypothesis and is based on maximum likelihood. There are two methods that have been used in classical factor analysis. The first, based on time-domain analysis, is the EM algorithm (e.g. Watson and Engle, 1983, and Quah and Sargent, 1993). The second is based on frequency-domain analysis (e.g., Sargent and Sims, 1977, and Geweke and Singleton, 1981).
3. Identification of the factors and the coefficients: The common factors, $\mathbf{u}_t$, are identified only up to an orthonormal rotation of dimension q. Expression (3.3) is observationally equivalent to

$$\mathbf{\Sigma}_x(\theta) = \mathbf{B}(e^{-i\theta})\mathbf{Q}(e^{-i\theta})\tilde{\mathbf{Q}}(e^{-i\theta})\tilde{\mathbf{B}}(e^{-i\theta}) + \mathbf{\Sigma}_\xi(\theta),$$

where $\mathbf{Q}(e^{-i\theta})\tilde{\mathbf{Q}}(e^{-i\theta}) = \mathbf{I}_q$. The identification of the factors and the coefficients has been analyzed by Geweke and Singleton (1981), who provide sufficient conditions for identification. As noticed by these authors, unitary transformation of $\mathbf{\Sigma}_u(\theta)$ is possible because the common factors are not dated. That is, if, in (3.2), we replace $\mathbf{u}_{j,t-s}$ by $\mathbf{u}_{j,t-s+\tau_j}$, for all s and t and $j = 1, \ldots, q$, the modified model will be observationally equivalent to (3.2), because a pure delay in the time domain is equivalent to a phase shift (rotation around the unit circle) in the frequency domain.

4. LARGE ECONOMIES: AN INFINITE DIMENSIONAL CROSS SECTION

When the number of time series is large (large n), a traditional factor analysis based on maximum likelihood estimation involves computational problems, because the number of parameters increases with n. The research question explored by the papers reviewed here is whether one can consistently extract the indexes as $n \to \infty$ and how n must be related to T as $(n, T) \to \infty$. The analysis for n going to infinity is what we call the analysis of large economies. The case

of large n is of great practical interest because many relevant business cycle questions involve the analysis of many variables for possibly many sectors, individuals, or regions.

The analysis for $n \to \infty$ helps not only for estimation, but also for the identification of the model. Interesting specifications imply a factor structure where the idiosyncratic components are not mutually orthogonal, but allow for some mild cross correlation. This is typically the case in asset pricing models (approximate factor structure), but also in macroeconomics when the object is to study the local effect of regional or sectoral specific shocks. For those specifications, if n is fixed, identification restrictions are not easily found. As we will see, they will come quite naturally when we analyze the model in population as $n \to \infty$.

Another point is that the analysis for $n \to \infty$ suggests a characterization of macroeconomic behavior. If, as n increases, $q < n$ sources of variations remain common to all variables, we can say that the economy is driven by q macroeconomic shocks and that this is the relevant stochastic dimension of macroeconomic models.

From now on we will think of the Xs as an infinite dimensional sequence indexed by n and study the properties of the model as n and T go to infinity.

Denote by $\{X_{it}; t \in \mathbb{Z}\}$ a double array of random variables; assume that the n one-dimensional series $\{X_{it}; t \in \mathbb{Z}\}$, $i = 1, \ldots, n$, have been observed over the period $t = 1, \ldots, T$; and write $\mathbf{X}_t^{(n)} = (X_{1t}, \ldots, X_{nt})'$ for the observation made at time t. The object of the study is the model

$$X_{it} = \chi_{it} + \xi_{it}, \quad t \in \mathbb{Z}, \tag{4.1}$$

where χ_{it}, the *common component*, can be represented as a dynamic linear combination

$$\chi_{it} = \sum_{j=1}^{q} b_{ij}(L) u_{jt}, \quad t \in \mathbb{Z}, \ i = 1, \ldots, n$$

of (a small number q of) unobservable *common shocks* $u_{1t}, \ldots, u_{qt}$, and ξ_{it}, the *idiosyncratic component*. The latter is orthogonal, at all leads and lags, to those common shocks, and hence to the dynamic space spanned by the common components.

We assume the following.

Assumption (B)

(B1). $\{\boldsymbol{u}_t = (u_{1t}, \ldots, u_{qt}); t \in \mathbb{Z}\}$ *is a zero-mean, normal white noise*;

(B2). *the coefficients of the nq filters* $b_{ij}(L) = \sum_{k=-\infty}^{\infty} b_{ijk} L^k$ *are square summable:* $\sum_{k=\infty}^{\infty} b_{ijk}^2 < \infty$;

(B3). $\{\boldsymbol{\xi}_t^{(n)} = (\xi_{1t}, \ldots, \xi_{nt})'; t \in \mathbb{Z}\}$ *is a stationary process such that*

$$\mathrm{E}\big[\boldsymbol{\xi}_t^{(n)}\big] = \mathbf{0} \quad \text{and} \quad \mathrm{E}[\boldsymbol{\xi}_{it} u'_{jt}] = 0, \quad i = 1 \ldots, n, \quad j = 1 \ldots, q, \quad t, t' \in \mathbb{Z}.$$

The corresponding statistical model for the vector of the observables is

$$\begin{aligned} \mathbf{X}_t^{(n)} &= \mathbf{B}^{(n)}(L)\mathbf{u}_t + \boldsymbol{\xi}_t^{(n)} \\ &= \boldsymbol{\chi}_t^{(n)} + \boldsymbol{\xi}_t^{(n)}, \end{aligned} \tag{4.2}$$

where $\mathbf{B}^{(n)}(L) = (\mathbf{b}_1^{(n)}(L), \ldots, \mathbf{b}_q^{(n)}(L))'$ is the $n \times q$ matrix with i column $\mathbf{b}_i^{(n)}(L) = (b_{1i}^{(n)}(L), \ldots, b_{ni}^{(n)}(L))'$ and $\boldsymbol{\chi}_t^{(n)} = (\chi_{1t}, \ldots, \chi_{nt})'$.

Assumption (B) implies stationarity of $\boldsymbol{X}_t^{(n)}$ for any n. This can accommodate processes that are stationary after some transformation such as differencing or deterministic detrending.

Denote by $\boldsymbol{\Sigma}^{(n)}(\theta)$ the spectral density of the observations and assume that its entries, $\sigma_{ij}(\theta)$, are bounded in modulus. Model (4.2) implies, as (3.2), that $\boldsymbol{\Sigma}^{(n)}(\theta)$ can be written as the sum of the spectral density of the χ s, $\boldsymbol{\Sigma}_{\chi}^{(n)}(\theta)$, which has reduced rank $q < n$, and the spectral density of the ξ s, $\boldsymbol{\Sigma}_{\xi}^{(n)}(\theta)$, which has rank n. However, because we have not imposed conditions on cross-sectional orthogonality of the idiosyncratic components, without further assumptions the model is not well specified.

The questions that have been discussed in the "large economies" literature are the same as those analyzed in the traditional factor models literature, but the analysis in population is for $n \to \infty$ whereas the statistical analysis establishes convergence for $(n, T) \to \infty$. This survey will review results on five issues.

1. *Identification:* Under what conditions (for $n \to \infty$) on the variance–covariance structure of the data is the common component identifiable?
2. *Representation:* Under what conditions (for $n \to \infty$) does a factor representation (4.2) exist?
3. *Estimation:* Find an (n, T) consistent estimator of the common component and derive relative rates of convergence for n and T.
4. *Forecasting:* Find an (n, T) consistent estimator of the minimum mean square linear forecast.
5. *Identification:* of $\mathbf{u}_t$ and the $\mathbf{b}_i^{(n)}$ coefficients (for $n \to \infty$).

4.1. Identification of the Components

Without cross-sectional orthogonality of the ξ_{it}, χ_{it} is not only nonobserved, but also not identified. Forni, Hallin, Lippi, and Reichlin (2000) have defined conditions on the spectral density matrix of the Xs under which the components are identified as n goes to infinity. The asymptotic identification conditions are preconditions to develop an estimator for the component that is consistent for n and T going to infinity.

The concept of asymptotic identification was first developed by Chamberlain (1983) and Chamberlain and Rothschild (1983) for a static factor model. Here I develop the analysis for the more general dynamic model and then state the static result as a special case.

The essential assumption for identification is as follows:

Assumption (A). *The nonzero q eigenvalues of $\mathbf{\Sigma}_{\chi}^{(n)}(\theta)$ diverge, whereas all eigenvalues of $\mathbf{\Sigma}_{\xi}^{(n)}(\theta)$ remain bounded, almost everywhere in $[-\pi, \pi]$, as $n \to \infty$.*

Assumption (A) is clearly satisfied if the ξ_{it} are mutually orthogonal at any lead and lag (and have uniformly bounded spectral densities) as in model (3.2), but is more general as it allows, so to speak, for a limited amount of dynamic cross correlation. Bounded eigenvalues of the spectral density matrix of the ξ imply that idiosyncratic causes of variation, although possibly shared by many (even all) units, have their effects concentrated on a finite number of units, and tending to zero as i tends to infinity. For example, the assumption is fulfilled if $\text{var}(\xi_{it}) = 1$, $\text{cov}(\xi_{it}, \xi_{i+1t}) = \rho \neq 0$, while $\text{cov}(\xi_{it}, \xi_{i+ht}) = 0$ for $h > 1$.

Similarly, the assumption of unbounded eigenvalues of the spectral density of the χs guarantees a minimum amount of cross correlation between the common components. With a slight oversimplification, this implies that each u_{jt} is present in infinitely many cross-sectional units with nondecreasing importance.

Model (4.2) under Assumption (A) is the *generalized dynamic factor model* of Forni et al. (2000) and Forni and Lippi (2001). This model is very general in the dynamic specification and flexible in the assumption on the idiosyncratic component; from now on, when refering to model (4.2) we will mean (4.2), under Assumptions (A) and (B).

Identification of the common component is shown in two steps. First, observe that by a law of large number argument, q "appropriately chosen" linear combinations of the Xs become increasingly collinear with the common factor as $n \to \infty$. This implies that q averages of the observations converge to the dynamic space spanned by the q common factors. As a result, and this is the second step, by projecting the Xs onto the present, past, and future of these averages, we converge to the common component, and we identify it uniquely (nonredundancy).

In Forni et al. (2000), it is shown that the first q dynamic ($q < n$) principal components of the Xs are "appropriate aggregates." The latter are a generalization of the well-known static concept (see Appendix B for details). They are q processes z_{jt}, $j = 1, \ldots, q$, linear combinations of the leads and lags of the variables in $\mathbf{X}_t$, that is,

$$z_{jt} = \underline{\mathbf{p}}_j(L)\mathbf{X}_t, \quad j = 1, \ldots, q,$$

where L is the lag operator and $\underline{\mathbf{p}}_j^{(n)}(L)$ is a $1 \times n$ row vector of two-sided filters that has been normalized in such a way that $\underline{\mathbf{p}}_j^{(n)}(L)\underline{\mathbf{p}}_k^{(n)}(F)' = 0$ for $h \neq k$ and $\underline{\mathbf{p}}_j^{(n)}(L)\underline{\mathbf{p}}_j^{(n)}(F)' = 1$, where the prime denotes transposition and $F = L^{-1}$.

As shown by Brillinger (1981), the projection of $\mathbf{X}_t$ on the present, past, and future of the first q (population) dynamic principal components provides, for given n, the best approximation (in mean square) to $\mathbf{X}_t$ by means of q linear

combinations of the leads and lags of the observations.[1] The projection can be written as

$$\begin{aligned}\chi_{it} &= [\underline{\tilde{p}}_{1,i}(L)\underline{\mathbf{p}}_1(L) + \underline{\tilde{p}}_{2,i}(L)\underline{\mathbf{p}}_2(L) + \cdots + \underline{\tilde{p}}_{q,i}(L)\underline{\mathbf{p}}_q(L)]\mathbf{X}_t \\ &= \underline{\mathbf{K}}_i(L)\mathbf{X}_t, \end{aligned} \tag{4.3}$$

where the k coefficient of $\underline{\mathbf{p}}_j(L)$ is

$$p_{jk} = \frac{1}{2\pi}\int_{-\pi}^{\pi} \mathbf{p}_j(\theta)e^{ik\theta}d\theta, \quad j = 1, \ldots, q,$$

with $\mathbf{p}_j(\theta)$ being the jq-row eigenvector of $\mathbf{\Sigma}(\theta)$ corresponding to the eigenvalue $\lambda_j(\theta)$. The coefficients of the projections, $\underline{\tilde{p}}_{j,i}(L)$, are the i elements of the vector $\underline{\tilde{\mathbf{p}}}_j(L)$, which is the conjugate transpose of $\underline{\mathbf{p}}_j^{(n)}(L)$, which we can also write as $\underline{\mathbf{p}}_j^{(n)}(F)'$.

Let us now index projection (4.3) by n. Forni et al. (2000) show that, as $n \to \infty$, the $\chi_{it}^{(n)}$ converges in mean square to the common component χ_{it} of the generalized dynamic factor model, (4.1).

RESULT FHLR1 (Forni et al., 2000). *Under Assumptions* (B) *and* (A), *we have*

$$\lim_{n\to\infty} \chi_{it}^{(n)} = \chi_{it}$$

in mean square for any i *and* t.

Because $\chi_{it}^{(n)}$ depends only on $\mathbf{X}_t^{(n)}$, Result FHLR1 has the immediate implication that the components χ_{it} and ξ_{it} are identified. Moreover, it is shown that representation (4.2) is *nonredundant*; that is, no other representation fulfilling Assumptions (A) and (B) is possible with a smaller number of factors. Notice that so far we have been assuming $T = \infty$, so $\chi_{it}^{(n)}$ is not an estimator, but a population quantity. Result FHLR1, however, is a basic building block for proving consistency to the common component of the corresponding empirical projection (see Section 4.3).

4.2. Representation

Unbounded eigenvalues of the spectral density of the common component and bounded eigenvalues of the spectral density of the idiosyncratic component, Assumption (A), imply that the first q eigenvalues of the observable spectral density of the observations are unbounded whereas the last $n - q$ are bounded (see Forni et al., 2000). Let us then make the following assumption.

[1] It is worth noting that the projection solving the maximization problem is unique, whereas the principal components themselves are not.

Assumption (A*). *The first q eigenvalues of* $\mathbf{\Sigma}^{(n)}(\theta)$ *diverge as* $n \to \infty$ *almost everywhere in* $[-\pi, \pi]$, *that is,* $\lim_{n\to\infty} \lambda_q^{(n)}(\theta) = \infty, \theta-$ *a.e., whereas all other eigenvalues are bounded as* $n \to \infty$: *there exists a* $\Lambda > 0$ *such that* $\lambda_{q+1}^{(n)}(\theta) \leq \Lambda$ *almost everywhere in* $[-\pi, \pi]$.

We have seen that Result FHLR1 in Forni et al. (2000) establishes that if Assumption (A) and therefore Assumption (A*) are satisfied, the generalized dynamic factor model, that is, representation (4.2) under Assumptions (A) and (B), exists. Forni and Lippi (2001) complete the representation theory for the generalized dynamic model by establishing an "if and only if" result.

RESULT FL (Forni and Lippi, 2001). *The double sequence* $\{X_{it}; t \in \mathbb{Z}\}$ *is a q-generalized dynamic factor model if and only if Assumption* (A*) *is satisfied.*

A consequence of this result is that evidence that, for some q, Assumption (A*) holds becomes evidence both that the series follows a generalized dynamic factor model and that the number of factors is q.

Result FL generalizes to the dynamic case, the result shown in Chamberlain (1983) and Chamberlain and Rothschild (1983) for the static approximate factor model:

$$X_i = c_{i1}v_1 + c_{i2}v_2 + \cdots + c_{iq}v_q + \rho_i. \tag{4.4}$$

Model (4.4) has no time dimension and is "isomorphic" to model (4.2) under the assumption that $b_{ij}(L)$ is trivial (does not contain leads or lags) and $\boldsymbol{\xi}_n$ is a white noise process. If this is the case, the spectral density of $\mathbf{X}^{(n)}$, its eigenvalues, and eigenvectors do not depend on θ, and all coincide with the variance–covariance matrix of $\mathbf{X}^{(n)}$, its eigenvalues, and eigenvectors, respectively, which are the tools employed in Chamberlain and Rothschild's analysis.

They show the following:

RESULT CR (Chamberlain, 1983, and Chamberlain and Rothschild, 1983). *If the vector* $\boldsymbol{X}_t^{(n)}$ *is a white noise for any n, that is, if the matrix* $\mathbf{\Sigma}^{(n)}(\theta)$ *and its eigenvalues are constant as functions of* θ, *then a q-factor representation exists if and only if the first q eigenvalues of the spectral density of the Xs are unbounded as a function of n, while the last* $n - q$ *are bounded.*

Results FL and CR establish a firm link between principal components and factor analysis.

As we will see, static analysis is sometimes used to treat the dynamic case. Assuming finite moving average representation for the common component, lagged factors can be treated as additional factors in a static setup such as (4.4). To appreciate the difference between this approach and the fully dynamic approach, consider, for instance,

$$X_{it} = u_t + \alpha_i u_{t-1} + \xi_{it}, \tag{4.5}$$

with $1 \geq \alpha_i \geq 1/2$ and $\boldsymbol{\xi}^{(n)}$ orthonormal white noise. Defining $v_{1t} = u_t$ and $v_{2t} = u_{t-1}$, model (4.5) is isomorphic to (4.4) with $q = 2$. As a consequence, the first two eigenvalues of the variance-covariance matrix of $\mathbf{X}^{(n)}$ diverge, whereas the third is bounded. However, an analysis of the eigenvalues of the variance–covariance matrix of (4.5) does not allow distinction between (4.5) and

$$X_{it} = v_{1t} + \alpha_i v_{2t} + \rho_{it}, \tag{4.6}$$

where v_{1t} and v_{2t} are orthogonal at any leads and lags. In contrast, an analysis of dynamic eigenvalues of the spectral density matrix gives the following.

1. Model (4.6) has constant spectral density. Therefore, dynamic and static analyses coincide. The first two eigenvalues diverge, whereas the third one is bounded.
2. The first eigenvalue of the spectral density matrix of (4.5) is not smaller than
$$\left\| \sum_{i=1}^{n} (1 + \alpha_i e^{-i\theta}) \right\|^2 = |n(1 + \bar{\alpha}_n e^{-i\theta})|^2,$$
where $\bar{\alpha}_n = \sum_{i=1}^{n} \alpha_i / n$, and therefore diverges for any $\theta \in [-\pi, \pi]$. The second dynamic eigenvalue is uniformly bounded.

Moreover, as soon as a model as simple as

$$X_{it} = \frac{1}{1 - \alpha_i L} u_t + \xi_{it}$$

is considered, a dynamic analysis reveals that the first eigenvalue diverges everywhere in $[-\pi, \pi]$, whereas the second one is uniformly bounded. By contrast, a static analysis leads to the conclusion that all eigenvalues of the variance–covariance matrix diverge. This is consistent with an infinite number of static factors.

4.3. Estimation and Convergence Rates

Result FHLR1 shows that the common component χ_{it} can be recovered asymptotically from the sequence $\underline{\mathbf{K}}_i^{(n)}(L)\mathbf{X}_t^{(n)}$. The filters $\underline{\mathbf{K}}_j^{(n)}(L)$ are obtained as functions of the spectral density matrices $\boldsymbol{\Sigma}^{(n)}(\theta)$. Now, in practice, the population spectral densities $\boldsymbol{\Sigma}^{(n)}(\theta)$ must be replaced by their empirical counterparts based on finite realizations of the form $\mathbf{X}_t^{(T,n)}$ (see Appendix C on details on estimation of the spectral density).

Let us assume the following.

Assumption (C). $\mathbf{X}_t^{(n)}$ *admits a linear representation of the form*

$$\mathbf{X}_t^{(n)} = \sum_{k=\infty}^{\infty} \boldsymbol{c}_k^{(n)} \mathbf{Z}_{t-k}^{(n)},$$

where $\{\mathbf{Z}_t^{(n)}; t \in \mathrm{Z}\}$ *is white noise with a nonsingular covariance matrix and finite fourth-order moments, and* $\sum_{k=-\infty}^{\infty}(\boldsymbol{c}_k^{(n)})_{ij}|k|^{1/2} < \infty$ *for* $i, j = 1, \ldots, n$.

Under Assumption (C), any periodogram-smoothing or lag-window estimator $\boldsymbol{\Sigma}^{(n,T)}(\theta)$ is a consistent (for $T \to \infty$) estimator of $\boldsymbol{\Sigma}^{(n)}(\theta)$ (see Brockwell and Davis, 1987; Appendix C provides details on the particular estimator used in the work reviewed here). Denote by $\lambda_j^{(n,T)}(\theta)$, $\underline{\mathbf{K}}_i^{(n,T)}(L)$ the estimated counterparts of $\lambda_j^{(n)}(\theta)$, $\underline{\mathbf{K}}_i^{(n)}(L)$, respectively, and put

$$\chi_{it}^{(n,T)} = \underline{\mathbf{K}}_i^{(n,T)}(L)\mathbf{X}_t^{(n)}(L).$$

Consider a truncation of this quantity (which is denoted by the same symbols so as not to burden the notation). The following result has been proved:

RESULT FHLR2 (Forni et al., 2000).

$$\lim \underline{\mathbf{K}}_i^{(n,T)}(L)\mathbf{X}_t^{(n)} = \chi_{it}$$

in probability for n *and* T *going to infinity at some rate.*

To prove Result FHLR2, write $|\boldsymbol{K}_i^{(n,T)}(L)\mathbf{X}_t^{(n)} - \chi_{it}|$ as the sum of $R_1^{(n,T)} = |\underline{\mathbf{K}}_i^{(n,T)}(L)\mathbf{X}_t^{(n)} - \underline{\mathbf{K}}_i^{(n)}(L)\mathbf{X}_t^{(n)}|$ and $R_2^{(n)} = |\underline{\mathbf{K}}_i^{(n)}(L)\mathbf{X}_t^{(n)} - \chi_{it}|$. $R_2^{(n)}$ depends on n only, so that convergence is guaranteed by result FHLR1; $R^{(n,T)}$ depends on both T and n and convergence can be proved by using Chebyshev's theorem.

Result FHLR2 is a simple consistency result, and it provides no consistency rates: it merely asserts the existence of paths in $\mathbb{N} \times \mathbb{N}$, of the form $\{(n, T(n)); n \in \mathbb{N}\}$, where $n \mapsto T(n)$ is monotone increasing and $T(n) \uparrow \infty$ as $n \to \infty$, such that the difference between $\chi_{it}^{(nT)}$ and χ_{it} goes to zero in probability as n and T tend to infinity along such paths. No information is provided about the form of these paths, that is, about $n \mapsto T(n)$.

A *consistency-with-rates* reinforcement of Result FHLR2 relates n, T, and the magnitude of the difference $|\chi_{it}^{(n,T)} - \chi_{it}|$ as n and T tend to infinity along certain paths $(n, T(n))$.

The difficulties in obtaining rates are that the estimator depends on the estimated spectral density matrix of the first n series, $\boldsymbol{\Sigma}^{(n,T)}(\theta)$. The latter is governed by a classical root-T rate, but the constants associated with such rates might depend on n, so that a given distance between estimated and population spectral density might require a T that increases faster than n.

For results to be obtained on consistency rates, additional assumptions on the model are needed. First, it suffices to impose an assumption of uniformity, as $n \to \infty$, on the variance of each of the common terms $b_{ij}(L)u_{jt}$. This can be interpreted as an assumption on the unobservable filters $b_{ij}(L)$: for instance, in the very simple model $X_{it} = u_t + \xi_{it}$, with independently identically distributed (iid.) idiosyncratic components ξ_{it}, the common shock u_t has a uniform impact on the n observed series. As a result, the (unique) diverging eigenvalue is $\lambda_1^{(n)}(\theta) = n + \sigma_\xi^2$, where σ_ξ^2 is the idiosyncratic variance,

with eigenvector $(n^{-1/2}, \ldots, n^{-1/2})$. Hence, $\Delta^{(n)} = \sqrt{n}$. No further restriction is needed on the idiosyncratic components. A second assumption is that the convergence of the estimated spectral density $\mathbf{\Sigma}^{(n,T)}(\theta)$ to the population spectral density $\mathbf{\Sigma}^{(n)}(\theta)$ is uniform with respect to n (both the rates of convergence and the constants associated with these rates are independent of n). $\mathbf{\Sigma}^{(n,T)}(\theta)$ is obtained from any smoothed periodogram method with bandwidth B_T that depends only on T. The latter should tend to zero as $T \to \infty$ – neither too fast nor too slow, if consistency of $\mathbf{\Sigma}^{(n,T)}(\theta)$ is to be achieved at appropriate rates – while $B_T \uparrow \infty$.

Now let $\delta^{(n)} > 0$ be an increasing sequence such that $\lim_{n\to\infty} \delta^{(n)} = \infty$.

Definition 4.1. *We say that* $\chi_{it}^{(n,T)}$ *converges* (A) *to the common factor space* $\mathcal{U}^{(n)}$[2] *at rate* $\delta^{(n)}$ *along the path* $(n, T(n))$ *if, for all* $\epsilon > 0$, *there exists a* $B_\epsilon > 0$ *and an* $N_\epsilon \in \mathbb{N}$ *such that*

$$\mathrm{P}[\delta^{(n)} |\chi_{it}^{(n,T(n))} - \mathrm{proj}_{\mathcal{U}^{(n)}} \chi_{it}^{(n,T(n))}| > B_\epsilon] < \epsilon$$

for all $n \geq N_\epsilon$.

(B) *to the common component at rate* $\delta^{(n)}$ *along the path* $(n, T(n))$ *if, for all* $\epsilon > 0$, *there exists a* $B_\epsilon > 0$ *and an* $N_\epsilon \in \mathbb{N}$ *such that*

$$\mathrm{P}\big[\delta^{(n)} |\chi_{it}^{(n,T(n))} - \chi_{it}| > B_\epsilon\big] < \epsilon$$

for all $n \geq N_\epsilon$.

We say that $\chi_{it}^{(nT)}$ achieves consistency rate $\delta^{(n)}$ along the path $(n, T(n))$.

RESULT FHLR3 (Forni, Hallin, Lippi, and Reichlin, 2002a). *Both rates of convergence are proved as long as* $T(n)$ *diverge, that is, at any relative rates for* n *and* T. *In the "classical" case of linearly diverging eigenvalues* ($\Delta^{(n)} = n^{1/2}$ *case*), *the rate of consistency, for convergence to the space of common components, depends only on* n *and is equal to* $n^{1/2}$. *The precision of the rate of convergence to the component, in contrast, also depends on* T *and the rate is of the form* $\min\{\sqrt{n}, \sqrt{B_T T}\}$.

The results can be illustrated by using the elementary example:

$$X_{it} = a_i u_t + \xi_{it}, \tag{4.7}$$

with ξ_{it} iid. and the assumption that $0 < a \leq a_i \leq b$ for any i. The estimator of, say, $a_1 u_t$ is an average of the X_{it-k}s, $i = 1, 2, \ldots, n$, thus an average of the u_{t-k} plus an average of the ξ_{it-k}. The second average tends to vanish in variance because the ξ are orthogonal, and this is quite independent of the speed at which T diverges. Thus the result that the rate at which the space

[2] $\mathcal{U}^{(n)}$ is the minimal closed subspace of $L_2(\Omega, \mathcal{F}, \mathrm{P})$ containing the first q-dynamic principal components.

spanned by the common shock is approached depends crucially on the rate at which n diverges, with the rate at which T diverges playing no role.

In contrast, in most interesting cases, a fast-diverging n does not help when the task is approaching the true common component $a_1 u_t$ because the estimated coefficients of the projections of the observations onto the common factors must also converge and the precision of their convergence depends on T. The cross-sectional dimension n efficiently contributes to the accuracy of the estimation, which, however, depends on the total number $nT(n)$ of observations.[3]

Note that (n, T)-consistent estimators (and rates) to the common component of factor models slightly different than (4.2) have been proved by Forni and Reichlin (1998) and Stock and Watson (1999).

Consider model (4.2) where Assumption (A) is replaced by the assumption of mutually orthogonal idiosyncratic elements. Then consider the minimal closed subspace of $L_2(\Omega, \mathcal{F}, \mathrm{P})$ containing q cross-sectional averages $\bar{\mathcal{U}}_n$ and the orthogonal projection

$$\bar{\chi}_{it}^{(n)} = \mathrm{proj}(x_{it} | \bar{\mathcal{U}}^{(n)}).$$

Forni and Reichlin (1998) have shown the following.

RESULT FR1 (Forni and Reichlin, 1998).

$$\lim \bar{\chi}_{it}^{(n)} = \chi_{it}$$

in probability for any i and t as $\min(n, T) \to \infty$.

The Forni–Reichlin estimator is the empirical projection on the present, past, and future of q cross-sectional averages. The consistency proof is constructed in a similar way to that of Result FHLR2. First, it is shown that under some conditions needed to avoid near singularities of the averages as n increases, the latter converge to the dynamic space spanned by the q common factors. Second, convergence to the common component is shown for the projections onto these aggregates. Consistency depends on the convergence properties of the coefficients of these projections, which depend on T. The Forni–Reichlin result tells us that consistency holds at any rate (independent of the relative speed at which n and T go to infinity), but does not derive explicit rates. Note than an advantage of using averages rather than dynamic principal components is that the latter are independent of the Xs and not estimated (as in the case of static or dynamic principal components). However, unless ad hoc assumptions are introduced, near singularity of the chosen averages for n growing, with the consequence of inaccurate estimation, cannot be excluded.

Let us now consider a different restriction of model (4.2), namely, assume that $b_{ij}(L)$ is of finite order m. Then model (4.2) can be written in its static form

[3] Note that, though its growth can be arbitrarily slow, the series length $T(n)$ has to go to infinity.

as an r-factor model where $r = (m + 1)q$:

$$\mathbf{X}_t^{(n)} = \mathbf{B}^{(n)}\mathbf{V}_t + \boldsymbol{\xi}_t^{(n)}. \tag{4.8}$$

Here $\mathbf{V}_t = (\mathbf{u}_t, \ldots, \mathbf{u}_{t-q})$ is $r \times 1$ and the ith row of $\mathbf{B}$ is $(b_{i10}, \ldots, b_{iqm})$.

In this framework, lags are treated as additional factors and (4.8) can be analyzed as a static factor model. For this model, and for fixed T, Connor and Korajczyk (1988, 1993) have shown that the first r static principal components of the Xs are (n) consistent to the factor space. They show that, when T is fixed, the problem of estimating the $n \times n$ variance–covariance matrix of the Xs can be turned into a $T \times T$ problem and compute static principal components of the $T \times T$ variance–covariance matrix of the Xs. Stock and Watson (1999) analyze the same estimator and show consistency for n and T going to infinity, and they provide results on relative rates of convergence.

Under suitable conditions on the covariance matrix of the ξ_{it}, which limit the cross-covariance of the idiosyncratic components, the results shown by Stock and Watson (1999) are as follows.

RESULT SW1 (Stock and Watson, 1999). *The static projection on the first r static principal components of the Xs converge in probability to the common component in (4.8) at any relative rate of n and T.*

RESULT SW2 (Stock and Watson, 1999). *The static projection on the first k static principal components of the Xs (where $k \geq r$) converges in probability to the space spanned by the r factors $\mathbf{V}_t$ for $n \gg T$.*

Result SW1 is a stronger result than SW2 because it shows convergence to the component rather than to the space. As in FR1, convergence is shown to occur independent of the relative rate of n and T, but, unlike in Result FHLR3, no explicit rates are derived. Result SW2, in contrast, is only for convergence to the space, but here explicit rates are derived. This is done for an estimator that is computed for a possibly misspecified number of factors and for a model that is more general than (4.8) in the sense that time-varying factor loadings are allowed for. Convergence rates, however, require the unpleasant condition $n \gg T$.

4.4. An Illustrative Example

To understand the intuition of Result FHLR1 and FHLR2, let us develop an example.

Let assume that we have a panel of normalized time series. Suppose that $q = 1$ and the filters $\mathbf{b}_i(L)$ appearing in Equation (4.2) are of the form L^{s_j}, with s_j equal to 0, 1, or 2. Thus we have a single common factor u_t; some of the variables load it with lag one – the coincident variables – some with lag zero – the leading variables – some with lag two – the lagging variables. Equation (4.2)

becomes

$$X_t^{(n)} = \begin{pmatrix} L^{s_1} \\ L^{s^2} \\ \vdots \\ L^{s_n} \end{pmatrix} u_t + \boldsymbol{\xi}_t^{(n)}.$$

Moreover, assume that the idiosyncratic components ξ_{jt}, $j = 1, \ldots, \infty$, are mutually orthogonal white noises, with the same variance σ^2, so that the spectral density of $\mathbf{X}_t^{(n)}$ is

$$\frac{1}{2\pi} \begin{pmatrix} e^{-is_1\theta} \\ e^{-is_2\theta} \\ \vdots \\ e^{-is_n\theta} \end{pmatrix} (e^{is_1\theta}\ e^{is_2\theta} \cdots\ e^{is_n\theta}) + \frac{\sigma^2}{2\pi}\mathbf{I}_n.$$

In this case, it can be easily verified that the larger eigenvalue is

$$\lambda_1^{(n)}(\theta) = n + \sigma^2, \tag{4.8}$$

and a valid corresponding row eigenvector is

$$\mathbf{p}_1^{(n)}(e^{-i\theta}) = \frac{1}{\sqrt{n}}(e^{is_1\theta}\ e^{is_2\theta} \cdots\ e^{is_n\theta}). \tag{4.9}$$

The related filter is[4]

$$\underline{\mathbf{p}}_1^{(n)}(L) = \frac{1}{\sqrt{n}}(F^{s_1}\ F^{s_2} \cdots\ F^{s_n}),$$

whereas the first principal component series is

$$\begin{aligned} z_{1t}^{(n)} &= \frac{1}{\sqrt{n}}(F^{s_1}\ F^{s_2} \cdots\ F^{s_n}) \begin{pmatrix} L^{s_1} \\ L^{s_2} \\ \vdots \\ L^{s_n} \end{pmatrix} u_t \\ &\quad + \frac{1}{\sqrt{n}}(F^{s_1}\ F^{s_2} \cdots\ F^{s_n})\,\boldsymbol{\xi}_t^{(n)} \\ &= \sqrt{n}u_t + \frac{1}{\sqrt{n}}\sum_{j=1}^{n} \xi_{jt+s_j}. \end{aligned}$$

[4] Note that, in this example, the idiosyncratic components play no role in the determination of $\underline{\mathbf{p}}_1^{(n)}(L)$, which would have been identical with zero idiosyncratic terms. This is due to the particular form that we have assumed here for the cross-covariance structure of the idiosyncratic components. However, the same property holds approximately for large n under Assumption (A), that is, the boundedness of the idiosyncratic eigenvalues of the spectral density matrix of the Xs.

Three observations are in order. First, the idiosyncratic part of the principal component vanishes with respect to the common part as n becomes larger and larger, so that the principal component itself becomes increasingly "collinear" with the common factor $\mathbf{u}_t$. In other words, if n is sufficiently large, the first principal component captures the information space spanned by the common shock (convergence to the space).

Second, the filter $\mathbf{p}_1^{(n)}(L)$ shifts the common components by multiplying each of the L^{s_j} precisely by F^{s_j}, so that time delays and time leads are eliminated and we end up by summing n times the same common shock u_t.

Third, the "estimated" common components, which coincide here with the χ_{jt}s appearing in Equation (4.2) (because of the simplification $T = \infty$) are

$$\chi_{jt}^{(n)} = \frac{1}{\sqrt{n}} L^{s_j} \left[\sqrt{n} u_t + \frac{1}{\sqrt{n}} \sum_{h=1}^{n} \xi_{ht+s_h} \right] = L^{s_j} u_t + \frac{1}{n} \sum_{h=1}^{n} \xi_{h(t+s_h-s_j)}$$
$$= \chi_{jt} + \frac{1}{n} \sum_{h=1}^{n} \xi_{h(t+s_h-s_j)}.$$

Therefore, when applying the filter $\mathbf{p}_1^{(n)}(F)'$, that is, projecting on the leads and lags of z_{1t} as in (4.3), the correct lags of the common components are restored and the leading or lagging nature of each variable emerges again. This is the convergence to the component result (Result FHLR1).

Let us comment on these three observations. The first result, "killing the idiosyncratic," could have been achieved by considering other linear combinations of the Xs than dynamic principal components. Any weighted average whose square coefficients tend to zero as $n \to \infty$ does the job (on this point, see Forni and Lippi, 2001). Forni and Reichlin (1998, 2001) have used simple cross-sectional averages and generalized static principal components; Chamberlain (1983) and Chamberlain and Rothschild (1983) have used static principal components to make the same point. Also notice that, for the more general case of a nonorthogonal idiosyncratic, Assumption (A) allows that $\underline{p}_{j,i}^{(n)}(F)' \underline{\mathbf{p}}_j^{(n)}(L) \boldsymbol{\xi}_t^{(n)}$ vanishes as n tends to infinity so that in the limit only the term $\underline{p}_{j,i}^{(n)}(F)' \underline{\mathbf{p}}_j^{(n)}(L) \boldsymbol{\chi}_t^{(n)}$ survives.

The second result, "realignment," is due to the fact that dynamic principal components weight leading and lagging variables appropriately. Notice that had we considered, as Forni and Reichlin (1998), cross-sectional averages as the aggregates onto which to project, the implied filter would have been $(1/n, \ldots, 1/n)'$ and all variables would have been weighted equally. Had we instead used the first static principal component, as in Stock and Watson (1999), the filter would have been $(1/\sqrt{n*}, \ldots, 1/\sqrt{n*})'$, where $n*$ is the number of the most numerous among the groups of leading, lagging, and coincident variables. This implies that we would have ended up selecting only the variables belonging to the most numerous group instead of picking variables among all groups of

coincident, leading, and lagging variables, possibly losing valuable information. To take into account the information on the less numerous groups, we would have had to consider more than one static principal component, possibly many. This poses a potential problem of efficiency and affects the approximation error, which is likely to increase as the number of aggregates increases because it is inversely related to the size of the aggregates (on this point, see Forni, Hallin, Lippi, and Reichlin, 2002b).

The third result, "reestablishment of correct leads and lags," is what shows that for convergence to the component – which is a stronger result than convergence to the space – we need not only to find appropriate aggregates, but also to project on the present, past, and future. Result FHLR1 shows that this ensures that we will capture the entire dynamic space spanned by the q factors. Projecting on static principal components does not ensure the result, as shown by the example $\chi_{it} = u_{t-i}$.

4.5. One-Sided Estimation and Forecasting

In a factor model, multivariate information can be exploited for forecasting the common component, whereas the idiosyncratic, being mildly cross correlated, can be reasonably well predicted by means of traditional univariate methods (or based on low-dimension models such as VARs). Therefore, the forecast of an element of the panel can be derived as the sum of the forecast of the common component, where we exploit multivariate information, and the forecast of the idiosyncratic component, where multivariate information can be disregarded. The common component being of reduced stochastic dimension, its forecast can be expressed as a projection on the span of few appropriately constructed aggregates. The methods proposed by Stock and Watson (1999) and Forni et al. (2002b) are both based on this general idea. Notice that in the forecasting context, multivariate information is valuable provided that the panel contains many leading variables with respect to the variable to be forecasted, because leading variables, loosely speaking, can help predict coincident and lagging ones.

To introduce the general idea, let us consider the following example:

$$(1-L)\mathbf{X}_t^{(4)} = \begin{pmatrix} 1 \\ L \\ -1 \\ -L \end{pmatrix} u_t + \boldsymbol{\xi}_t^{(4)},$$

where we have a panel of four variables driven by one common factor that loads with different coefficients on different elements of the cross section: the first variable is procyclical and leading, the second is procyclical and lagging, the third is countercyclical and leading, and the fourth is countercyclical and lagging.

This model can also be written in its static version as

$$(1 - L)\mathbf{X}_t^{(4)} = \begin{pmatrix} u_{1t} \\ u_{2t} \\ -u_{1t} \\ -u_{2t} \end{pmatrix} + \boldsymbol{\xi}_t^{(4)},$$

where $u_{1t} = u_t$ and $u_{2t} = u_{t-1}$.

In the example, the number of dynamic factors is $q = 1$ while the number of static factors is $r = 2$. Using the argument developed in Section 4.3, we find that this implies that one aggregate should converge to the space spanned by u_t and two aggregates should converge to the space spanned by u_{1t} and u_{2t}. Consistent estimation of the common component should be obtained with a projection onto the span of one dynamic aggregate or through a projection onto the span of two static aggregates. As discussed in Section 4.3, the former method is what is suggested by Forni et al. (2000), and the latter is what is suggested by Stock and Watson (1999). These ideas can be applied to the forecasting problem.

Defining $\mathcal{G}(\mathbf{u}, T)$ as the span of the common factors $u_{ht}, h = 1, \ldots, q$, we see that the optimal linear forecast of the common component (i.e., the minimum square error forecast) is

$$\phi_{i,T+h} = \text{Proj}[\chi_{i,T+h} | \mathcal{G}(\mathbf{u}, T)]. \tag{4.10}$$

The optimal h-step ahead forecast of $X_{i,T+h}$ is the sum of this projection and the optimal linear forecast of the idiosyncratic. Given the small cross-sectional correlation of the elements of the idiosyncratic components, the latter can be expressed as a univariate autoregressive model, so that the optimal linear forecast of the variable i is

$$X_{i,T+h} = \phi_{i,T+h} + \sum_{j=1}^{p} \alpha_i \xi_{i,t-j}.$$

The problem is to obtain an (n, T)-consistent estimate of the projection, that is, a consistent estimate of $\mathcal{G}(\mathbf{u}, T)$ and of the covariances $\boldsymbol{\Gamma}_{\chi,k}$ and $\boldsymbol{\Gamma}_{\xi,k}$.

Let us consider model (4.2) with the restriction of finite lag structure for the factors, as in model (4.8). The method of aggregation based on the static principal component advocated by Stock and Watson can be applied to the forecast problem. The Stock and Watson (1999) result follows directly from Result SW1 discussed in Section 4.3. Because the static rank of $\mathcal{G}(\mathbf{u}, T)$ is $r = q(m + 1)$ (equal to 2 in the example), they can estimate $\mathcal{G}(\mathbf{u}, T)$ by the first r static principal components of the Xs. Calling the vector of the first r static principal components $\mathbf{Z}_t$, we have

$$\phi_{i,T+h}^{(n),T} = \left(\boldsymbol{\Gamma}_{\chi,h}^{(n),T} \mathbf{Z}^{(n),T} \left(\mathbf{Z}^{(n),T\prime} \boldsymbol{\Gamma}_{\chi,0}^{(n),T} \mathbf{Z}^{(n),T}\right)^{-1} \left(\boldsymbol{X}^{(n),T}\right)_i\right.$$

and the following.

RESULT SW1 BIS (Stock and Watson, 1999).

$$\phi_{i,T+h}^{(n),T} \to \phi_{i,T+h}$$

in probability for $(n, T) \to \infty$ *at any relative rate of* n *and* T.

Notice that the estimated projection is obtained by static aggregation of the observations, that is, as a linear combination of current values of the observations. Under finite lag structure, the result can easily be generalized to the stacked case where the weights are obtained from the covariance matrix of stacked observations.

As we have seen, the Stock and Watson (1999) method is potentially less efficient than the method based on dynamic principal components advocated by Forni et al. (2000) (two regressors instead of one for the example herein). However, estimation results from Forni et al. (2000) cannot easily be extended to obtain consistent forecasts because the dynamic principal component method produces an estimator to the common component, which is a two-sided filter of the observations. Although their method has the advantage of exploiting the dynamic structure of the data and needs relatively few dynamic regressors (aggregates) to approximate the space spanned by the common factors, two-sidedness is obviously an unpleasant characteristic for forecasting.

Forni et al. (2002b) propose a refinement of the original procedure that retains the advantages of the dynamic approach while obtaining a consistent estimate of the optimal forecast as a one-sided filter of the observations. The method consists of two steps. In the first step, they follow earlier work (Forni et al., 2000) and obtain the cross-covariances for common and idiosyncratic components at all leads and lags from the inverse Fourier transforms of the estimated spectral density matrices. Let us define them as $\mathbf{\Gamma}_{\chi,h}^{(n),T}$ and $\mathbf{\Gamma}_{\xi,h}^{(n),T}$, respectively. In the second step, they use these estimates to obtain the r contemporaneous linear combinations of the observations with the smallest idiosyncratic–common variance ratio. The resulting aggregates can be obtained as the solution of a *generalized principal component* problem.

More precisely, they compute the generalized eigenvalues μ_j, that is, the n complex numbers solving $\det(\mathbf{\Gamma}_{\chi,0}^{(n),T} - z\mathbf{\Gamma}_{\xi,0}^{(n),T}) = 0$ and the corresponding generalized eigenvectors $\mathbf{V}_j^{(n),T}$, $j = 1, \ldots, n$, that is, the vectors satisfying

$$\mathbf{V}_j^{(n),T}\mathbf{\Gamma}_{\chi,0}^{(n),T} = \mu_j \mathbf{V}_j^{(n),T}\mathbf{\Gamma}_{\xi,0}^{(n),T},$$

and the normalizing condition

$$\begin{aligned} \boldsymbol{V}_j^{(n),T}\mathbf{\Gamma}_{\xi,0}^{(n),T}\mathbf{V}_i^{(n),T\prime} &= 0 \quad \text{for } j \neq i, \\ &= 1 \quad \text{for } j = i. \end{aligned}$$

Then they order the eigenvalues in descending order and take the eigenvectors corresponding to the largest r eigenvalues. The estimated static factors are the

generalized principal components

$$v_{jt}^{(n),T} = \mathbf{V}_j^{(n),T\prime}\mathbf{X}_t^{(n),T}, \quad j = 1, \ldots, r.$$

The generalized principal components have a useful efficiency property: they are the linear combinations of the X_{jt}s having the smallest idiosyncratic–common variance ratio (for a proof, see Forni et al., 2002b).

By using the generalized principal components and the covariances estimated in the first step, Forni et al. obtain an alternative estimate of $\mathcal{G}(\boldsymbol{u}, T)$. The estimated projection is

$$\phi_{i,T+h}^{(n),T} = \left(\boldsymbol{\Gamma}_h^{(n),\chi}\mathbf{V}^{(n),T}\left(\mathbf{V}^{(n),T\prime}\boldsymbol{\Gamma}_0^{(n),\chi}\mathbf{V}^{(n),T}\right)^{-1}(\mathbf{X}^{(n),T})_i\right.$$

and the following.

RESULT FHLR4 (Forni et al., 2002b).

$$\phi_{i,T+h}^{(n),T} \to \phi_{i,T+h}$$

in probability for $(n, T) \to \infty$ *at a suitable rate of* n *and* T.

Notice that the same method can be used to reestimate the within-sample common component, thus improving the accuracy of the estimator based on the first step. These projections do not involve future observations and hence do not suffer the end-of-sample problems of the earlier method (Forni et al., 2000).

Both the Stock and Watson estimators and those of Forni et al. are linear combinations of present and past observations, but the weighting schemes used are different. Theoretically, they both provide a consistent forecast. Empirical relative performances, however, are difficult to establish a priori. Indeed, though the weighting scheme of Forni et al., being tailored with the purpose of minimizing the impact of the idiosyncratic in the aggregate, should perform better in approximating the common factor space, relative performance depends in a very complicated way on the underlying model, on T, and on n. Forni et al., (2002b) report simulation results comparing small sample performances of the two methods.

4.6. Identification of the Factors and Coefficients: Factor Models and Structural VARs

Model (4.2) has no structural meaning, and the factors $\mathbf{u}_t$ are identified up to (dynamic) orthogonal rotations (see Section 3). Forni and Reichlin (1998) have made the point that, as far as identification is concerned, there is a close similarity between factor models and structural VAR (SVAR) models because, in both cases, we face the same kind of rotational indeterminacy. They have also noticed that, as long as n is finite, in the factor model the problem is

to identify the joint moments of the factors, the factors loadings, not the factors themselves. The common shocks are inherently unobservable and cannot be identified. In contrast, when n is infinite, the factors can be identified and the similarity with SVARs is even closer. Because the idiosyncratic disturbances die out in the aggregate, for the aggregate variables the factor model collapses in a VAR model so that we can use the same identification strategy used for SVARs to identify the shocks and therefore the parameters of the factor model.

The Forni and Reichlin argument is as follows. Assume that the $\mathbf{b}_j(L)$ filters in (4.2) are one sided and let $\mathbf{z}_t = \mathbf{D}(L)\mathbf{u}_t$ denote a fundamental infinite moving average representation of a $q \times 1$ vector of aggregates. Fundamentalness implies that

$$\overline{\text{span}}(\mathbf{z}_{t-k}, k \geq 0) \supseteq \overline{\text{span}}(\mathbf{u}_{t-k}, k \geq 0),$$

that is, that $\mathbf{u}_t$ can be recovered from a projection onto the present and past of $\mathbf{z}_t$.

The vector $\mathbf{z}_t$ could be any weighted averages of the Xs whose square coefficients tend to zero as $n \to \infty$ (simple cross-sectional averages or principal components). This condition ensures $\mathbf{z}_t$ to be the n-mean square limit of $\mathbf{z}_t^{(n)}$ (see Forni and Lippi, 2001). The Wold representation of the vector of the aggregates can be recovered as

$$\mathbf{z}_t = \mathbf{D}(L)\mathbf{D}(0)^{-1}\mathbf{v}_t,$$

where $\mathbf{v}_t = \mathbf{D}(0)\mathbf{u}_t$.

By inverting the Wold representation and reordering terms, one can obtain the VAR

$$\mathbf{z}_t = \mathcal{D}(L)\mathbf{z}_{t-1} + \mathbf{v}_t.$$

Having obtained an estimate for $\mathbf{v}_t$ and $\mathcal{D}(L)$, the authors suggest identifying $\mathbf{D}(L)$ as in the SVAR literature. This implies the following steps. Orthonormality of the shocks is imposed by switching to representation

$$\mathbf{z}_t = (\mathbf{D}(L)\mathbf{D}(0)^{-1}\mathbf{U}^{-1})\mathbf{U}\mathbf{v}_t = \hat{\mathbf{D}}(L)\hat{\mathbf{u}}_t,$$

where $\mathbf{U}$ is the upper triangular matrix such that $\mathbf{U}\boldsymbol{\Sigma}_v\mathbf{U}^{(-1)} = \mathbf{I}_q$. Any alternative fundamental moving average representation corresponding to a given structural model can be obtained from an orthonormal rotation of $\hat{\mathbf{D}}(L)\hat{\mathbf{u}}_t$; that is, it will be of the following form:

$$\mathbf{z}_t = \mathbf{C}(L)\mathbf{e}_t,$$

with $\boldsymbol{\Sigma}_{\boldsymbol{e}} = \mathbf{I}_q$, $\mathbf{e}_t = \mathbf{R}'\hat{\mathbf{u}}_t$, and $\mathbf{C}(L) = \hat{\mathbf{D}}(L)\mathbf{R}$, where $\mathbf{R}$ is a constant orthonormal matrix such that $\mathbf{R}\mathbf{R}' = \mathbf{I}_q$.

Hence, for the model to be identified, it suffices to select a $q \times q$ static rotation $\mathbf{R}$.

Forni and Reichlin (1996) establish the following result.

RESULT FR2 (Forni and Reichlin, 1996). *If* $\mathbf{u}_t$ *is fundamental for* $\mathbf{z}_t$, *then none of the* X_{it} *Granger causes* $\mathbf{z}_t$.

The intuition of this result is that, if none of the individual variables are leading with respect to the vector of aggregates, which spans the same space as the common component, then the common shocks $\mathbf{u}_t$ must belong to the space spanned by the present and past of the aggregates. Therefore, fundamentalness, unlike for SVARs, is testable.

Forni, Lippi, and Reichlin (2002) explore this point further and show (a) that conditions under which the common factors can be recovered from the present and past of the observations ($\mathbf{u}_t$ fundamental for X_{it}, for all i) are easily met in factor models; and (b) how to derive the fundamental representation and identify the loadings $\mathbf{b}_i(L)$ and the factors $\mathbf{u}_t$.

Result (a) can be understood by the fact that, in factor models, the problem is to identify few common shocks from information on many variables. This implies that fundamentalness does not require conditions on each n rows of $\mathbf{B}(L)$ in (4.2),[5] but just the existence of an $n \times q$ one-sided filter $\mathbf{C}(L)$ such that

$$\mathbf{C}(L)'\mathbf{B}(L) = \mathbf{I}_q.$$

This is indeed a very mild condition. If a $q \times q$ invertible submatrix of $\mathbf{B}(L)$ exists, there is no problem. If not, the left inverse could still exist. Consider the following example. Assume that $q = 1$ and that each line of $\mathbf{B}(L)$ is non-fundamental moving average of order one, for example, $b_i(L) = 1 - a_i L$ with $a_i > 1$ so that none of them is invertible. Nonentheless, if $a_i \neq a_j$,

$$\frac{(1 - a_i L)a_j - (1 - a_j L)a_i}{a_j - a_i} = 1.$$

Therefore we can set $c_h(L) = 0$ for $h \neq i, j$; $c_i(L) = a_j/(a_j - a_i)$; and $c_j(L) = a_i/(a_j - a_i)$. Hence the only case in which we have nonfundamentalness is

$$\mathbf{B}(L) = \mathbf{B}(0)\left(1 - \frac{1}{a}L\right),$$

and the fundamental noise is

$$w_t = \frac{1 - aL}{1 - (1/a)L}u_t.$$

Once fundamentalness is established, the problem of identification is reduced to the selection of a static rotation of dimension q. Notice that, in this framework, the number of required restrictions depends on q and not on n. What matters is

[5] This is under the assumption that (each element of) $\mathbf{b}_i(L)$, $i = 1, \ldots, \infty$, in (4.1) is unilateral toward the past.

the number of "common shocks," not the number of variables. This contrasts with the SVAR framework, where the rotation problem has dimension n and the number of required restrictions for just identification grows with the number of included variables.

This point strongly advocates for the use of factor models for structural macroeconomic analysis. Typically, economic theory does not provide clear guidance on what variables to consider, and we know from the SVAR literature that results on impulse response functions are not robust with respect to the conditioning variables. Not only the rotation is arbitrary: the choice of variables is as well. In the factor framework, we can, in principle, include all potentially useful information and then extract the common shocks q that are likely to be a small number as compared with n. Once the relevant dimension is known, identification is a $q \times q$ problem. Moreover, conditions for fundamentalness are easier to find than in SVARs.[6]

5. ECONOMETRIC PROBLEMS

Identification of the number of dynamic factors q and of the number of static factors r. Criteria for the choice of q in the large economies literature are heuristic. In the discussion of Section 4 we have assumed that q, the number of nonredundant common factors, is known. In practice, of course, q is not predetermined and also has to be selected from the data. Result FL can be used to this end, because it links the number of factors in (4.2) to the eigenvalues of the spectral density matrix of $\mathbf{X}^{(n)}$: precisely, if the number of factors is q and $\boldsymbol{\xi}$ is idiosyncratic, then the first q dynamic eigenvalues of $\boldsymbol{\Sigma}^{(n)}(\theta)$ diverge a.e. in $[-\pi, \pi]$, whereas the $(q+1)$-th one is uniformly bounded.

Indeed, no formal testing procedure can be expected for selecting the number q of factors in finite sample situations. Even letting $T \to \infty$ does not help much. The definition of the idiosyncratic component indeed is of an asymptotic nature, where asymptotics are taken as $n \to \infty$, and there is no way a slowly diverging sequence (divergence, under the model, can be arbitrarily slow) can be told from an eventually bounded sequence (for which the bound can be arbitrarily large). Practitioners thus have to rely on a heuristic inspection of the eigenvalues against the number of series n.

More precisely, if T observations are available for a large number n of variables x_{it}, the spectral density matrices $\boldsymbol{\Sigma}_r^T(\theta)$, $r \leq n$, can be estimated, and the resulting empirical dynamic eigenvalues $\lambda_{rj}^T(\theta)$ computed for a grid of frequencies. The following two features of the eigenvalues computed from $\boldsymbol{\Sigma}_r^T(\theta)$, $r = 1, \ldots, n$, should be considered as reasonable evidence that the data have been generated by (4.1), with q factors, and that $\boldsymbol{\xi}$ is idiosyncratic:

[6] For an analysis of this problem, see Lippi and Reichlin (1993).

1. The average over θ of the first q empirical eigenvalues diverges, whereas the average of the $(q+1)$-th one is relatively stable.
2. Taking $r = n$, we find there is a substantial gap between the variance explained by the qth principal component and the variance explained by the $(q+1)$-th one. A preassigned minimum, such as 5 percent, for the explained variance could be used as a practical criterion for the determination of the number of common factors to retain.

Connor and Korajczyk (1993), in contrast, developed a formal test for the number of factors in a static factor model under sequential limit asymptotics; that is, n converges to infinity with a fixed T and then let T converge to infinity. Their test is valid under assumptions on cross-sectional stationarity (mixing conditions). The problem with this strategy is that a cross-sectional mixing condition is hard to justify given that the cross section, unlike time, has no natural ordering.

Stock and Watson (1999) show that, assuming $n, T \to \infty$ with $\sqrt{n}/T \to \infty$, a modification to the BIC criterion can be used to select the number of factors optimal for forecasting a single series. As observed by Bai and Ng (2002), their criterion is restrictive not only because it requires $n \gg T$, but also because there can be factors that are pervasive for a set of data and yet have no predictive ability for an individual data series. Thus, their rule may not be appropriate outside the forecasting framework.

Bai and Ng (2002) consider the static factor model and set up the determination of the number of factors as a model selection problem. The proposed criteria depend on the usual trade-off between good fit and parsimony.

Liska (2001) extends the Bai and Ng criterion to the dynamic factor model case.

Asymptotic Distribution. Results described so far establish consistency of the estimators. Bai (2001) has started developing an inferential theory for static factor models of large dimension and derived asymptotic distribution (and rates of convergence) of the estimated factors, factor loadings, and the common components.

Testing. The Generalized Dynamic Factor Model (GDFM) provides a parsimonious way for us to take into account the large information set assumption when estimating shocks and propagation mechanisms. As we have seen in Section 4.6, we can identify few shocks from many variables by imposing a minimal set of economically motivated restrictions. Because the shocks to be identified are less numerous than the variables we are conditioning on, with a set of minimal restrictions we can easily obtain overidentification restrictions. Giannone (2002) has developed testing procedures for long-run restrictions in this context.

The paper deals with hypothesis testing problems involving the loading coefficients of the common shocks and proposes Wald-type test statistics that

converge to a chi-square distribution as both the number of series and the number of observations go to infinity.

Bai and Ng (2001b) explore unit root testing in large factor models and propose a new approach. The idea is to exploit the information contained in a large panel of time series to perform a common–idiosyncratic decomposition and then to test stationarity of the common and idiosyncratic components separately. Although unit root tests are imprecise when the common components and the idiosyncratic have different orders of integration, direct testing on the two components is found to be more accurate in simulations.

6. APPLICATIONS

As a way to illustrate some of the potential uses of the methodology discussed for revealing features of business cycle behavior, three applications are described on different data sets.

6.1. The U. S. Manufacturing Sector

The paper by Forni and Reichlin (1998) investigates the behavior of output and productivity for 450 manufacturing sectors in the United States from 1958 to 1986.

- What is the degree of cross-sectional commonality in the time-series behavior of sectoral output and productivity?

A first answer to this question can be given by assessing the relative importance of the common and idiosyncratic components. An overall measure of fit is the ratio of the sum of the variances of the common components to the sum of the total variances of the variables. This, which is the weighted mean of the sectoral R^2 with weights proportional to the total variances, gives us a percentage of 41 percent for output and of 29 percent for productivity.

- Is the degree of commonality stronger at business cycle frequencies?

Overall variance ratios are not sufficiently informative about the role of idiosyncratic shocks for business cycle fluctuations. For this we must look at the distribution across frequencies of the variances of the common and sectoral components. This is captured by the sum of the spectra for the common and the idiosyncratic component (Figure 2.1).

Notice that, for both variables, whereas the common component has a typical business cycle shape with a peak corresponding to a period of just over four years, the bulk of the variance for the idiosyncratic component is spread equally over frequencies, suggesting that the latter can be characterized as a white noise, probably capturing measurement error. We should conclude that the business cycle features in manufacturing are mostly explained by economywide shocks and that, although the sectoral dynamics is more sizeable than

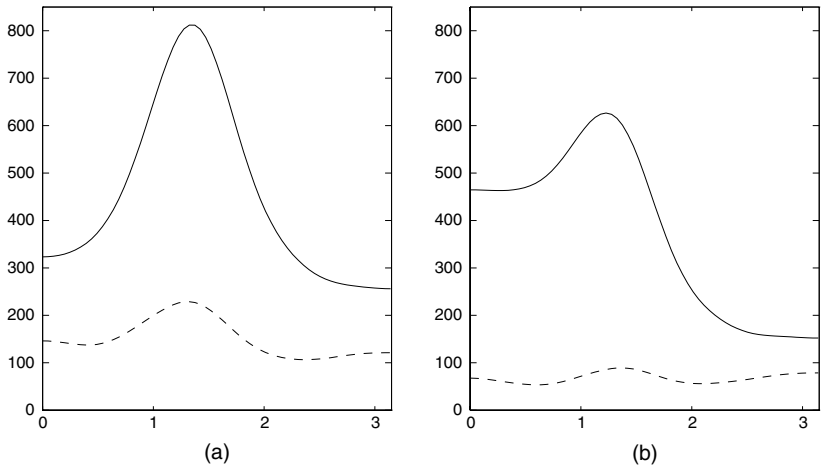


Figure 2.1. Sum of the spectra of the common and idiosyncratic components of (a) output and (b) productivity. Note: common component (solid line), idiosyncratic component (dashed line). Horizontal axis: frequencies $\theta \in [0, \pi]$, where $\theta = 2\pi/P$ and P is the cycle's period.

the economywide one, it cannot account for the cyclical behavior of output and productivity.

- Number of common shocks and their propagation over the 450 sectors of output and productivity.

The heuristic method proposed in the paper identifies two common shocks, one of which is identified as the technology shock (see the paper for details, as well as Section 4.5). To analyze the weight of the substitution or reallocation effects in the total variability of output and productivity, the authors look at the correlation structure of the impulse response functions associated with the two aggregate shocks. Let us call *substitution effects* the negative sectoral comovements generated by the aggregate shocks and *complementary effects* the positive sectoral comovements. The former can be measured by the absolute value of the sum of the negative cospectra $|\sum \sigma_{ij}(\theta)_-|$, whereas the latter can be measured by the absolute value of the sum of the positive cospectra $|\sum \sigma_{ij}(\theta)_+|$.

Figure 2.2 reports $|\sum \sigma_{ij}(\theta)_-|$ and $|\sum \sigma_{ij}(\theta)_+|$ for the technology shock and the nontechnology shock (the technology shock here is identified by selecting a particular orthonormal rotation of the moving average representation of the vector of the cross-sectional averages of output and productivity).

The figure illustrates nicely the business cycle features of our data: All the series of the sums of the positive cospectra have peaks at business cycle frequencies, while the series of the negative cospectra are rather flat. Moreover, the business cycle is partly real because the technology shock generates positive cospectra at a period of about four years.

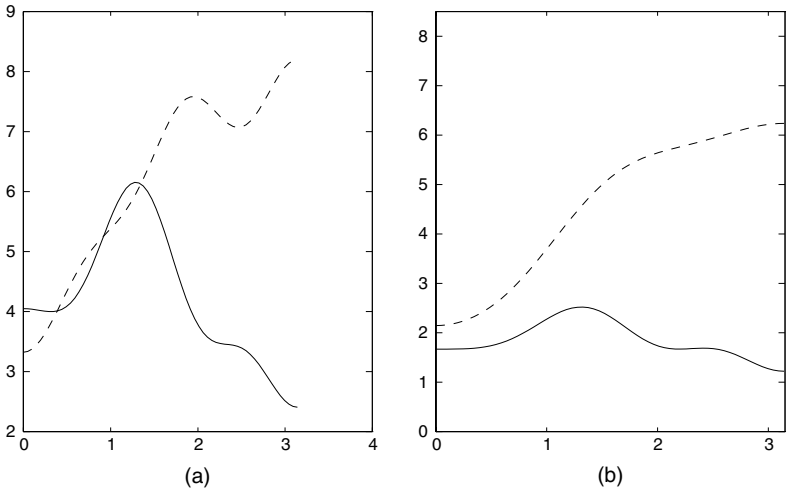


Figure 2.2. Absolute sum of positive (solid curves) and negative (dashed curves) cospectra. Note: technological component (a), nontechnological component (b). Horizontal axis: frequencies $\theta \in [0, \pi]$, where $\theta = 2\pi/P$ and P is the cycle's period.

6.2. Europe Versus the United States: Regions (Counties), Nations (States), and Aggregate Behavior

The study by Forni and Reichlin (2001) analyzes a data set on output growth for 138 regions of nine European countries and 3,075 counties for 48 U.S. states. The period considered, for most of the data set, is 1970–1993.

Let us denote with y_t^{ij} the growth rate of output for the ith region of nation j, expressed in deviation from the time-series mean. The authors assume

$$y_t^{ij} = E_t^{ij} + N_t^{ij} + \mathcal{L}_t^{ij} = a^{ij}(L)e_t + b^{ij}(L)n_t^j + c^{ij}(L)l_t^{ij}, \qquad (6.1)$$

for $j = 1, \ldots, J$ and $i = 1, \ldots, I^j$. Here E_t^{ij}, N_t^{ij}, and $\mathcal{L}_t^{ij}$ are the European component, the national component, and the local component, respectively; $a^{ij}(L)$, $b^{ij}(L)$, and $c^{ij}(L)$ are functions in the lag operator L. The European shock e_t, the national shocks n_t^j, and the local shocks l_t^{ij} are unobserved unit-variance white noises, mutually uncorrelated at all leads and lags.

The difference of this model with respect to the traditional dynamic factor model with orthogonal idiosyncratic components is that the factor n_t^j is neither common nor idiosyncratic. It is an intermediate-level factor, common for regions belonging to the same country but orthogonal across countries. The difference from the generalized factor model is that additional restrictions on the spectral density matrix of the observations are imposed.

- Regional, national, and overnational commonality.

For most European countries, the common European component explains the bulk of output variance. Exceptions are Greece, Portugal, and the UK, which

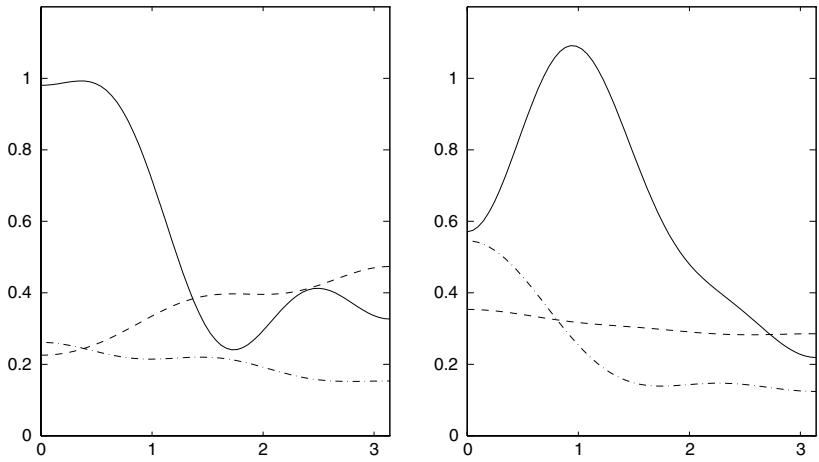


Figure 2.3. The spectral shape of the three components for Europe and the United States. Note: Horizontal axis: frequency; vertical axis: spectral density. Common component: solid line; National (state) component: dotted and dashed line; local component: dotted line. Europe (first data set, UK excluded) is on the left; United States (large and medium counties) is on the right.

have a large nation-specific volatility, well above 50 percent. Idiosyncratic regional variance is also sizeable (about 30 percent on average), while the nation-specific volatility is the smallest. Results for the United States are strikingly similar. The average size of the U.S.-wide component is similar to that of the European component when Greece, Portugal, and the UK are excluded, and the U.S. state component seems to be of the same order of magnitude as the national component in Europe. In both cases, global and local dynamics seem to prevail over national dynamics. This result is especially surprising for the European case. It indicates that pre–monetary union Europe already had a high degree of business cycle commonality and that nation-specific cycles do not account for much of the total output volatility.

- Cross-sectional dynamics.

Figure 2.3 shows the average spectra of the three components for Europe (UK excluded) and the United States (medium and large counties). Although, as we have seen, the variances of the output components are similar, the dynamic profiles are very different. There are two main differences between Europe and the United States. First, the European common component is very persistent, whereas the U.S.-wide component exhibits a typical business cycle shape, which peaks at a period of around six years. By looking at low frequencies, we see that the total long-run variance is similar in Europe and the United States, so that the uncertainty about the future income level at, say, ten or twenty years is nearly the same. However – and this is the second difference – whereas in the United States the main bulk of long-run fluctuations is state specific or local, in Europe the long-run variance is mainly common. This implies that European regions

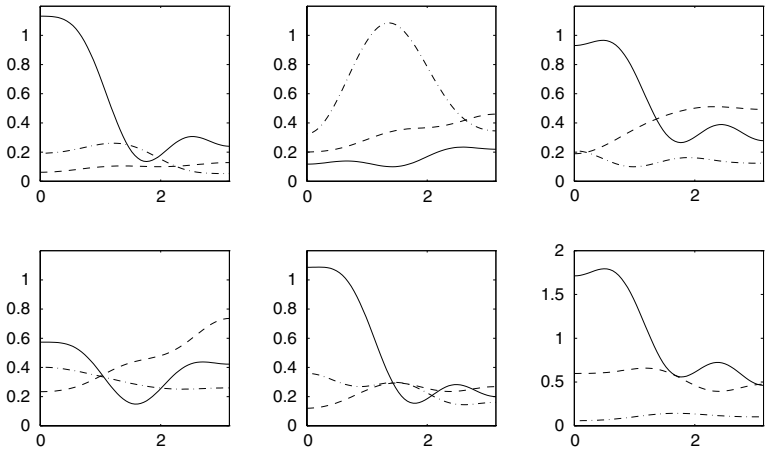


Figure 2.4. The spectral shape of the three components for six European countries. Note: Horizontal axis: frequency; vertical axis: spectral density. Common component: solid line; National (state) component: dotted and dashed line; local component: dotted line. From the left to the right: Germany, UK, France; Italy, Belgium, Netherlands (Groningen excluded). Netherlands has a different scale on the vertical axis.

have a larger long-run covariance, that is, larger cospectra near the vertical axis. In other words, European regions have a "common destiny" in the long run, more than U.S. counties. Notice that this does not imply that European countries have similar income levels, or that they are converging to the same income level. The result says only that persistent shocks are mainly common. Still, both drift and initial levels, which are not analyzed here, could be rather different, leading to permanent gaps, convergence, or even divergence.[7]

The analysis disaggregated by country (Figure 2.4) shows that the UK is the only European nation that, like the United States, has a typical business cycle. The other countries confirm the aggregate result of a large Europeanwide component, with most of the variance concentrated at low frequency (although Italy has a lot of high-frequency variation in the idiosyncratic).

- Synchronization and symmetry.

Before we can interpret these results on variance decomposition in terms of degree of integration, we must study the degree of synchronization and symmetry of the propagation mechanisms of the common shock. If the response functions of the common shock had different signs, the interpretation of a large common component at low frequency would imply that regions are diverging, exactly the opposite of what is concluded thus far.

[7] Also note that we are analyzing total income, as opposed to per capita income. Different dynamics of total income in the United States and Europe could be compensated for by migrations, leading to similar dynamics of per capita income.

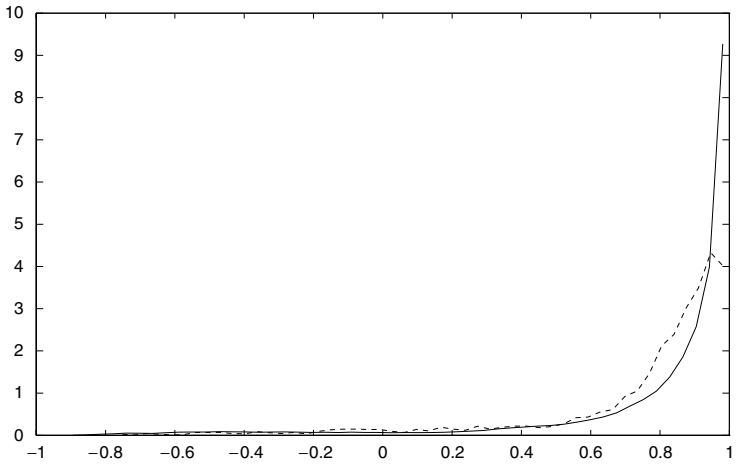


Figure 2.5. Density distribution of correlations among the common components. Note: Europe: dashed line; United States: solid line. Horizontal axis: correlation coefficients.

A rough measure of symmetry and synchronization is given by the regional distribution of the correlation coefficients. In Figure 2.5 we report the frequency density distribution of a grid of correlation intervals for European regions and U.S. medium and large counties. In both cases, we can see that most correlation coefficients are positive and large.[8] This confirms the interpretation of variance decomposition results given herein, even if we can notice that European regional dynamics is slightly more asymmetric than in the U.S. case.

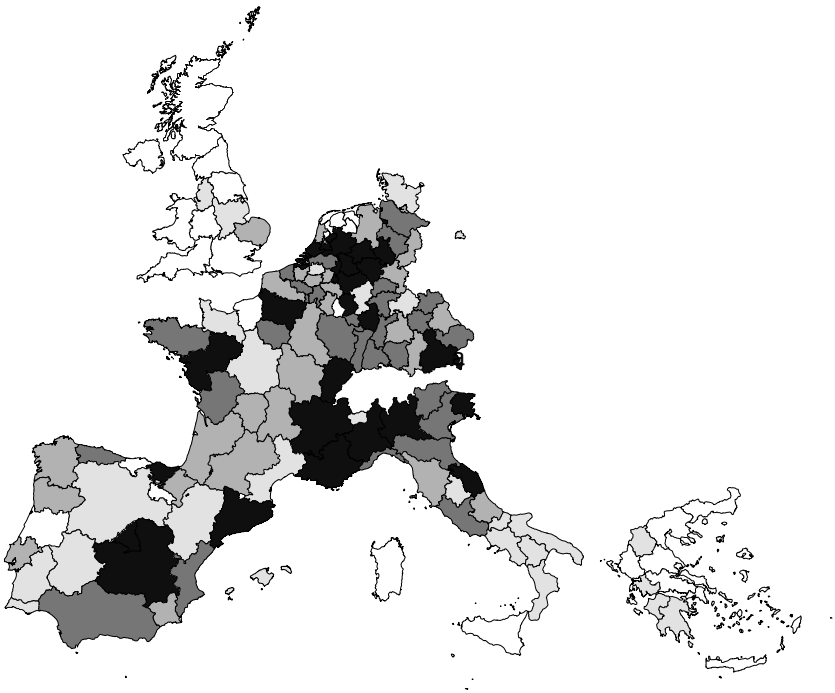
- Geography

Finally, to complete the picture on European integration, we want to ask the question of whether regions that are "more European" (larger relative variance of the European component) belong to a particular geographical area. Map 2.1 reports the geographical distribution of variance ratios between Europeanwide components and total variance. Light gray indicates a small European component; dark gray indicates a large European component.

Map 2.1 shows that a core made by the key countries France, Germany, and the Benelux does not exist. Dark and light spots are sparse, indicating that almost all countries are partly in and partly out. The only exceptions are Greece and the UK, which are clearly less integrated with the rest of Europe. In general, heterogeneity within nations seems at least as large as heterogeneity across nations.

In summary, European regions are already highly integrated and are expected to "move together" in the long run more than U.S. counties. The common

[8] The negative values are accounted for by Sicily, Sardinia, some UK regions, and Groningen, the Dutch outlier.



Map 2.1. Percentage of output variance explained by the European component. Note: Dark regions have a large European component. Limits for color changes are 0.23, 0.42, 0.58, 0.70.

shock exhibits high persistence in Europe and a typical business cycle shape in the United States. If we exclude Greece and the UK, Europe appears to be a continent of regions rather than nations, as far as output fluctuations are concerned.

6.3. EuroCOIN: A Real-Time Coincident Indicator for the Euroarea Business Cycle

The paper by Altissimo et al. (2001) develops a methodology for the construction of a monthly indicator of the Euroarea business cycle that is released every month without waiting for the current monthly publication of industrial production or current quarterly publication of the GDP. It refines the idea given by Forni et al. (2001a). The index is based on approximately 1,000 monthly time series for the six major countries of the Euroarea, Belgium, France, Germany, Italy, the Netherlands, and Spain, since 1987. It is now published by the Centre for European Policy Research (CEPR) on the 28th of every month (consult www.cepr.org).

The basic idea driving the construction of the coincident indicator is that the GDP is a good summary measure of economic activity. However, the GDP is

affected by errors and noise that disturb the true underlying signal. Such disturbances are measurement errors, local and sectoral shocks, and high-frequency, low-persistence movements. The procedure, based on the generalized dynamic factor model, is designed to clean the GDP of such disturbances.

Let the first variable in our panel be the European real GDP. Then EuroCOIN is defined as χ_1^C, that is, the cyclical, common component of the European GDP.

Why smooth the GDP by eliminating the short-run part of the common component? Monthly data are typically affected by large seasonal and higher-frequency sources of variation. Both economic agents and policy makers are not particularly interested in such high-fequency changes because of their transitory nature. Washing out temporary oscillations is necessary to unveil the true underlying long-lasting tendency of the economic activity.

Why clean the GDP of the idiosyncratic component? There are two reasons for this: eliminating measurement errors and producing a better signal for policy makers. First, national GDPs are not obtained by means of direct observation, but are at least in part the result of estimation procedures and therefore are affected by estimation errors. Moreover, data on GDP are provided quarterly by the statistical institutes; monthly figures can be obtained only by interpolating original data, which entails additional errors. Finally, the European GDP stems from an aggregation of data provided by heterogeneous sources, not all equally reliable and perfectly comparable. Summing up, the European GDP is affected by large measurement and estimation errors. Such errors are mainly idiosyncratic, because they are poorly correlated across different variables and independent from the common shocks. Second, the idiosyncratic component should capture both variable-specific shocks, such as shocks affecting, say, the output of a particular industrial sector, and local-specific shocks, such as, for instance, a natural disaster, having possibly large but geographically concentrated effects. Distinguishing between such shocks and common shocks, affecting all sectors and areas, can be useful for policy makers, who have to decide whether to carry out local and sectoral measures or common, Europeanwide interventions.

The full estimation procedure is in four steps.

The first step is to estimate the covariances of the unobservable components. Covariances are obtained by applying the inverse Fourier transform to the estimated spectral density (see Appendix D). The covariances for the cyclical and the noncyclical components χ_{jt}^C and χ_{jt}^S are obtained by applying such a transformation to the selected frequency band of the estimated spectra and cross-spectra (see Appendix D).

The second step is to estimate the static factors. In the second step, the authors compute an estimate of the static factors, following Forni et al. (2002b). By the term "static factors" we mean the $r = q(m + 1)$ variables appearing contemporaneously in representation (4.8), including the lagged u_t, so that, say, u_{1t} and u_{1t-1} are different static factors.

The third step is to estimate the cyclical common components. In the third step, they use contemporaneous, past, and future values of the static factors to

obtain an estimate of χ_{1t}^C, the cyclical component of the GDP. Precisely, they project χ_{1t}^C on $\mathbf{v}_{t-m}, \ldots, \mathbf{v}_{t+m}$, where $\mathbf{v}_t = (v_{1t}, \ldots, v_{rt})'$. The lag-window size m should increase with the sample size T, but at a slower rate. Consistency of such an estimator is ensured, for appropriate relative rates of m, T, and n, by the fact that (a) the projection of χ_{1t}^C on the first m leads and lags of χ_{1t} is consistent because of consistency of $\chi_{1t}^{(n),T}$ and the estimated covariances involved; and (b) χ_{1t} is a linear combination of the factors in $\mathbf{v}_t$, so that projecting on the factors cannot be worse than projecting on the common component itself.

Notice that this method is something like a multivariate version of the procedure given by Christiano and Fitzgerald (2001) to approximate the band-pass filter. Exploiting the superior information embedded in the cross-sectional dimension makes it possible to obtain a very good smoothing by using a very small window ($m = 1$). This has the important consequence that a timely and reliable end-of-sample estimation is obtained without having to revise the estimates for a long time (say, 12 months or more) after the first release, as with the univariate procedure. To get an intuition of the reason why they get good results with a narrow window, consider the extreme case $m = 0$. Clearly, with univariate prediction, one cannot get any smoothing at all. In contrast, the static factors will include, in general, both contemporaneous and past values of the common shocks and can therefore produce smooth linear combinations.

The fourth step is the end-of-the-sample unbalance. Finally, data become available with different delays. Had we to wait until the last updating arrives, we would be able to compute the indicator only with a delay of four or five months. The authors propose a procedure to handle this problem, which allows us to get provisional estimates by exploiting, for each series, the most updated information. Once the missing data become available, the final estimate is computed. This is why the indicator is subject to revision for a few months.

The procedure to handle this problem is the following. Let T be the last date for which the whole data set is available. Until T they estimate the static factors as just explained, that is, by taking the generalized principal components of the vector $\mathbf{X}_T^{(n),T}$. From T onward, they use the generalized principal components of a modified n-dimensional vector $\mathbf{Y}_T^{(n),T}$, which includes, for each process in the data set, only the last observed variable, in such a way as to exploit for each process the most recent information. Clearly, computation will involve the estimated covariance matrices of the common and the idiosyncratic component of $\mathbf{Y}_T^{(n),T}$ in place of those of $\mathbf{X}_T^{(n),T}$.

The idea is simply to shift the variables in such a way as to retain, for each one of them, only the most updated observation, and compute the generalized principal components for the realigned vector. In such a way one is able to get information on the factors u_{hT+j}, $h = 1, \ldots, q$, $j = 1, \ldots, w$, and to exploit it in prediction. The forecasts are then used to replace missing data and to get the forecasts of $\boldsymbol{\chi}_{T+h}$, $h > w$.

Figure 2.6 reports the EuroCOIN since 1987:1 plotted against monthly growth of GDP (both at a quarterly rate). The chart also marks the dating

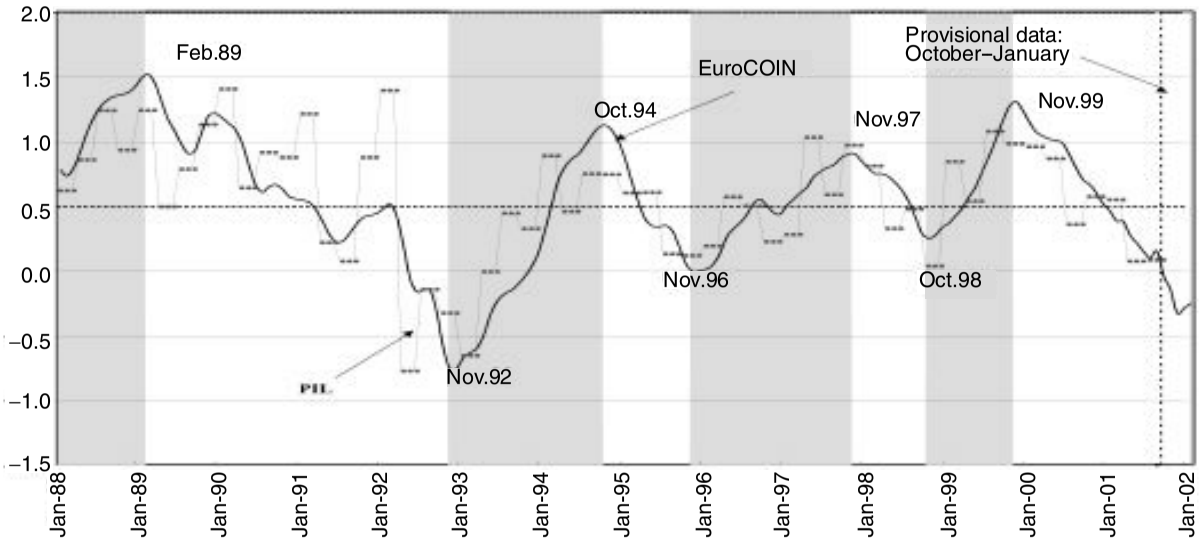


Figure 2.6. EuroCOIN and Euroarea GDP (quarterly rate of change).

of the European business cycle. Shaded areas are expansions, that is, periods of prolonged increasing growth. Notice that, as opposed to the classifical definition used by the National Bureau of Economic Research (NBER), which is in terms of business cycle levels, the definitions of recessions and expansions are in terms of rates of growth (growth cycle concept).

A by-product of the methodology is that one can track the contribution of different variables or blocks of variables (sectors or nations) to the aggregate result and evaluate leadership and commonality of different blocks. These "facts" are reported in the paper and a selection of them regularly published on the web page of the CEPR. Obviously, at the end of the sample, the estimates are based on a reduced number of variables where the weight of financial variables and survey data, which are released with a minimum delay, is particularly large.

6.4. Other Applications

Other applications of the model to macroeconomic and financial variables are the following.

1. *Business cycles*: This is the construction of a monthly index of core inflation for the European area (Cristadoro et al., 2001).
2. *Monetary policy*: This is the propagation of monetary policy shocks across European countries (Sala, 2001) and the analysis of systematic and unsystematic monetary policy in the United States (Giannone, Reichlin, and Sala, 2002; and Bernanke and Boivin, 2001).
3. *Links between financial and real variables*: This is the role of financial variables in predicting output and inflation in the European area (Forni et al., 2001b and D'Agostino, 2001), and the use of factor models for measuring the degree of real and financial markets integration (Emiris, 2001).
4. *Exchange rate dynamics*: This is the cross-sectional predictability of exchange rates (Rodrigues, 2002).

ACKNOWLEDGMENTS

This paper was prepared for the World Congress of the Econometric Society, August, 2000. Visit www.dynfactors.org

This survey draws heavily from joint work with Mario Forni, Marc Hallin, and Marco Lippi, whom I thank. I also thank my exceptional research assistants, Jeorges Rodrigues and Luca Sala.

APPENDIX A: THE SPECTRAL REPRESENTATION

Any stationary variable can be represented as the integral of waves of different frequency, each having a random amplitude; this is the so-called spectral

representation,

$$x_t = \int_{-\pi}^{\pi} e^{i\theta t} dZ(\theta),$$

where $dZ(\theta)$ is an "orthogonal increment process" such that $\text{cov}(dZ(\theta), dZ(\lambda)) = 0$ for $\lambda \neq \theta$.

The spectral density function of a stationary process is defined as

$$\sigma(\theta) = (1/2\pi) \sum_{h=-\infty}^{\infty} e^{ih\theta} \gamma_h, \quad -\infty < \theta < \infty,$$

where γ_h is the covariance function. Conversely, the covariance function can be obtained from the spectral density function by using Fourier techniques.

In the multivariate case, there are analogous definitions. The spectral density matrix of a stationary vector process is defined as

$$\mathbf{\Sigma}(\theta) = (1/2\pi) \sum_{h=-\infty}^{\infty} e^{ih\theta} \mathbf{\Gamma}_h, \quad -\infty < \theta < \infty,$$

where $\mathbf{\Gamma}_k$ is the covariance function. The latter can be obtained by Fourier inversion as

$$\mathbf{\Gamma}_k = \int_{-\pi}^{\pi} e^{i\theta h} \mathbf{\Sigma}(\theta) d\theta$$

(see, e.g., Brockwell and Davis, Chapter 4).

APPENDIX B: ESTIMATION OF THE SPECTRAL DENSITY

The estimation of spectral density in the various papers by Forni, Hallin, Lippi, and Reichlin is constructed by using a Bartlett lag-window estimator of size $M = M(T)$.

The sample covariance matrix $\mathbf{\Gamma}_k^{(T,n)}$ of $\mathbf{X}_t^{(n)}$ and $\mathbf{X}_{t-k}^{(n)}$ is computed for $k = 0, 1, \ldots, M$. Then, the authors compute the $(2M + 1)$ points discrete Fourier transform of the truncated two-sided sequence $\mathbf{\Gamma}_{-M}^{(T,n)}, \ldots, \mathbf{\Gamma}_0^{(T,n)}, \ldots, \mathbf{\Gamma}_M^{(T,n)}$, where $\mathbf{\Gamma}_{-k}^{(n,T)} = \mathbf{\Gamma}_k^{(n,T)\prime}$. More precisely, they compute

$$\mathbf{\Sigma}^{(T,n)}(\theta_h) = \sum_{k=-M}^{M} \mathbf{\Gamma}_k^{(T,n)} \omega_k e^{-ik\theta_h}, \tag{B.1}$$

where

$$\theta_h = 2\pi h/(2M + 1), \quad h = 0, 1, \ldots, 2M,$$

and $\omega_k = 1 - \{(|k|)/[(M + 1)]\}$ are the weights corresponding to the Bartlett lag window of size $M = M(T)$. The choice of M represents the trade-off between small bias (large M) and small variance (small M). To ensure consistency

of $\mathbf{\Sigma}^{(T,n)}(\theta)$, which is required for the validity of Result FHLR2, the condition $M(T) \to \infty$ and $M(T)/T \to 0$ as $T \to \infty$ must be fulfilled.

In Forni et al. (2000) a fixed rule $M = \text{round}(\sqrt{T}/4)$ is used and shown to perform well in simulations.

APPENDIX C: DYNAMIC PRINCIPAL COMPONENTS AND THE FILTER OF FORNI ET AL. (2000)

The dynamic principal component decompositions are obtained as in Brillinger (1981). From now on, let us drop the superscript (n), T for notational simplicity. For each frequency of the grid, the eigenvalues and eigenvectors of $\mathbf{\Sigma}(\theta)$ are computed. By ordering the eigenvalues in descending order for each frequency and collecting values corresponding to different frequencies, we obtain the eigenvalue and eigenvector functions $\lambda_j(\theta)$ and $\mathbf{p}_j(\theta)$, $j = 1, \ldots, n$. The function $\lambda_j(\theta)$ can be interpreted as the (sample) spectral density of the jth principal component series, and, in analogy with the standard static principal component analysis, the ratio

$$p_j = \int_{-\pi}^{\pi} \lambda_j(\theta) d\theta \Bigg/ \sum_{j=1}^{n} \int_{-\pi}^{\pi} \lambda_j(\theta) d\theta$$

represents the contribution of the jth principal component series to the total variance in the system. Letting $\mathbf{\Lambda}_q(\theta)$ be the diagonal matrix having on the diagonal $\lambda_1(\theta), \ldots, \lambda_q(\theta)$ and letting $\mathbf{P}(\theta)$ be the $(n \times q)$ matrix $(\mathbf{p}_1(\theta) \cdots \mathbf{p}_q(\theta))$, we see that the estimate of the spectral density matrix of the vector of the common components $\chi_t = (\chi_{1t} \cdots \chi_{nt})'$ is given by

$$\mathbf{\Sigma}_\chi(\theta) = \mathbf{P}(\theta)\mathbf{\Lambda}(\theta)\tilde{\boldsymbol{P}}(\theta), \tag{C.1}$$

where the tilde denotes conjugation and transposition.

The filter is computed from the first q eigenvectors $\mathbf{p}_j(\theta_h)$, $j = 1, 2, \ldots, q$, of $\mathbf{\Sigma}(\theta_h)$, for $h = 0, 1, \ldots, 2M$. For $h = 0, 1, \ldots, 2M$, Forni et al. (2000) construct

$$\mathbf{K}_i(\theta_h) = \tilde{p}_{1,i}(\theta_h)\mathbf{p}_1(\theta_h) + \cdots + \tilde{p}_{q,i}(\theta_h)\mathbf{p}_q(\theta_h).$$

The proposed estimator of the filter $\underline{\mathbf{K}}_j^{(n),T}(L)$, $j = 1, 2, \ldots, q$, is obtained by the inverse discrete Fourier transform of the vector

$$\left(\mathbf{K}_i^{(T,n)}(\theta_0), \ldots, \mathbf{K}_i^{(T,n)}(\theta_{2J})\right),$$

that is, by the computation of

$$\underline{\mathbf{K}}_{i,k}^{(T,n)} = \frac{1}{2J+1} \sum_{h=0}^{2J} \mathbf{K}_i^{(T,n)}(\theta_h) e^{ik\theta_h}$$

for $k = -J, \ldots, J$. The estimator of the filter is given by

$$\underline{\mathbf{K}}_i^{(T,n)}(L) = \sum_{k=-J}^{J} \underline{\mathbf{K}}_{i,k}^{(T,n)} L^k. \tag{C.2}$$

APPENDIX D: ESTIMATES OF MEDIUM AND LONG-RUN COVARIANCES OF THE COMPONENTS

Starting from the estimated spectral density matrix, estimates of the covariance matrices of χ_t at different leads and lags can be obtained by using the inverse discrete Fourier transform, that is,

$$\mathbf{\Gamma}_{\chi,k} = \left(\frac{2\pi}{101}\right) \sum_{h=-50}^{50} \mathbf{\Sigma}_{\chi}(\theta_h) e^{i\theta_h k},$$

where the superscript (n), T has been dropped for notational simplicity and will be dropped from now on.

Estimates of the covariance matrices of the medium and long-run component $\chi_t^C = (\chi_{1t}^C, \ldots, \chi_{nt}^C)'$ can be computed by applying the inverse transform to the frequency band of interest, that is, $[-\theta^*, \theta^*]$, where $\theta^* = 2\pi/24$, so that all periodicities shorter than one year are cut off (this is the cut-off point in Altissimo et al., 2001). Precisely, letting $\mathbf{\Gamma}_{\chi^C}(k) = E(\chi_t^C \chi_{t-k}^{C\prime})$, we see that the corresponding estimate will be

$$\mathbf{\Gamma}_{\chi^L}(k) = \left(\frac{2\pi}{2H+1}\right) \sum_{h=-H}^{H} \Sigma_{\chi}(\theta_h) e^{i\theta_h k},$$

where H is defined by the conditions $\theta_H \leq 2\pi/24$ and $\theta_{H+1} > 2\pi/24$.

References

Altissimo, F., A. Bassanetti, R. Cristadoro, M. Forni, M. Lippi, L. Reichlin, and G. Veronese (2001), "A Real Time Coincident Indicator of the Euro Area Business Cycle," Working Paper, CEPR.

Bai, J. (2001), "Inference on Factor Models of Large Dimension," mimeo, John Hopkins University.

Bai, J. and S. Ng (2001a), "A New Look at Panel Testing of Stationarity and the PPP Hypothesis," October mimeo, John Hopkins University.

Bai, J. and S. Ng (2001b), "A Panic Attack on Unit Roots and Cointegration," December mimeo, John Hopkins University.

Bai, J. and S. Ng (2002), "Determining the Number of Factors in Approximate Factor Models," *Econometrica*, 70(1), 191–221.

Bernanke, B. and J. Boivin, "Monetary Policy in a Data-Rich Enviroment," *Journal of Monetary Economics* (forthcoming).

Brillinger, D. R. (1981), *Time Series Data Analysis and Theory.* New York: Holt, Rinehart and Winston.

Brockwell, P. J. and R. A. Davis (1987), *Time Series: Theory and Methods*. New York: Springer-Verlag.

Burns, A. F. and W. C. Mitchell (1946), *Measuring Business Cycles.* New York: NBER.

Chamberlain, G. (1983), "Funds, Factors, and Diversification in Arbitrage Pricing Models," *Econometrica*, 51, 1281–1304.

Chamberlain, G. and M. Rothschild (1983), "Arbitrage, Factor Structure, and Mean-Variance Analysis in Large Asset Markets," *Econometrica*, 51, 1305–1324.

Connor, G. and R. A. Korajczyk (1988), "Risk and Return in an Equilibrium APT. Application of a New Test Methodology," *Journal of Financial Economics*, 21, 255–289.

Connor, G. and R. A. Korajczyk (1993), "A Test for the Number of Factors in an Approximate Factor Model," *Journal of Finance*, 48(4), 1263–1291.

Christiano, L. J. and T. J. Fitzgerald (2001), "The Band Pass Filter," July, NBER no. Working Paper W7257.

Cristadoro, R. M. Forni, L. Reichlin, and G. Veronese (2001), "A Core Inflation Index for the Euro Area," CEPR, Working Paper.

D'Agostino, A. (2002, January), "Understanding the Leading-Lagging Structure of Sectoral Returns," mimeo, ECARES, Universite Libre de Bruxelles.

Emiris, M. (2001, May), "Measuring Capital Market Integration," mimeo, ECARES, Universite Libre de Bruxelles.

Favero, C. and M. Marcellino (2001, December), "Large Data Sets, Small Models, and Monetary Policy in Europe," CEPR, Working Paper.

Forni, M., M. Hallin, M. Lippi, and L. Reichlin (2000), "The Generalized Dynamic Factor Model: Identification and Estimation," *The Review of Economics and Statistics*, 82, 540–554.

Forni, M., M. Hallin, M. Lippi, and L. Reichlin (2001a), "Coincident and Leading Indicators for the Euro Area," *The Economic Journal*, 111, 62–85.

Forni, M., M. Hallin, M. Lippi, and L. Reichlin (2001b), "Do Financial Variables Help Forecasting Inflation and Real Activity in the Euro Area?" CEPR, Working Paper.

Forni, M., M. Hallin, M. Lippi, and L. Reichlin (2002a), "The Generalized Dynamic Factor Model: Consistency and Convergence Rates," *Journal of Econometrics*, 82, 540–554.

Forni, M., M. Hallin, M. Lippi, and L. Reichlin (2002b), "The Generalized Dynamic Factor Model: One-Sided Estimation and Forecasting," January, mimeo, ECARES, Universite Libre de Bruxelles.

Forni, M. and M. Lippi (1997), *Aggregation and the Microfoundations of Dynamic Macroeconomics*. Oxford: Oxford University Press.

Forni, M. and M. Lippi (2001), "The Generalized Factor Model: Representation Theory," *Econometric Theory*, 17, 1113–1141.

Forni, M. M. Lippi, and L. Reichlin (2002), "Opening the Black Box: Identifying Shocks and Propagation Mechanisms in VARs and Factor Models," January, mimeo, ECARES, Universite Libre de Bruxelles.

Forni, M. and L. Reichlin (1996), "Dynamic Common Factors in Large Cross Sections," *Empirical Economics*, 21, 27–42.

Forni, M. and L. Reichlin (1998), "Let's Get Real: A Factor Analytic Approach to Disaggregated Business Cycle Dynamics," *Review of Economic Studies*, 65, 453–473.

Forni, M. and L. Reichlin (2001), "National Policies and Local Economies," *European Economic Review*, January, 82, 540–554.

Geweke, J. (1977), "The Dynamic Factor Analysis of Economic Time Series," in *Latent Variables in Socio-economic Models*, (ed. by D. J. Aigner and A. S. Goldberger), Amsterdam: North-Holland.

Geweke, J. and K. J. Singleton (1981), "Maximum Likelihood 'Confirmatory' Factor Analysis of Economic Time Series," *International Economic Review*, 22, 37–54.

Giannone, D. (2002), "Testing for Linear Restrictions with Large Panels of Time Series," January, mimeo, ECARES, Universite Libre de Bruxelles.

Giannone, D., L. Reichlin, and L. Sala (2002), "Tracking Greenspan: Systematic and Unsystematic Monetary Policy Revisited," January, mimeo, ECARES, Universite Libre de Bruxelles.

Lippi, M. and L. Reichlin (1993), "The Dynamic Effects of Aggregate Demand and Supply Disturbances: Comment," *American Economic Review*, 83, 644–652.

Liska, R. (2001), "Dynamic Factor Analysis: The Number of Factors and Related Issues," December, mimeo, ECARES, Universite Libre de Bruxelles.

Marcellino, M. J. H. Stock, and M. W. Watson (2002), "Macroeconomic Forecasting in the Euro Area: Country Specific Versus Area-Wide Information," *European Economic Review* (forthcoming).

Quah, D. and T. J. Sargent (1993), "A Dynamic Index Model for Large Cross Sections," in *Business Cycles, Indicators, and Forecasting* (ed. by J. H. Stock and M. W. Watson), Chicago: NBER and University of Chicago Press.

Rodrigues, J. (2002), "Common and Idiosyncratic Shocks to the Dynamics of Exchange Rate Volatility," February, mimeo, ECARES, Universite Libre de Bruxelles.

Sala, L. (2001, October), "Monetary Transmission in the Euro Area: A Factor Model Approach," mimeo, ECARES, Universite Libre de Bruxelles.

Sargent, T. J. and C. A. Sims (1977), "Business Cycle Modelling without Pretending to Have Too Much a *Priori* Economic Theory," in *New Methods in Business Research*, (ed. by C. A. Sims), Minneapolis: Federal Reserve Bank of Minneapolis.

Stock, J. H. and M. W. Watson (1989), "New Indexes of Coincident and Leading Economic Indicators," *NBER Macroeconomics Annual*, 1989, 351–394.

Stock, J. H. and M. W. Watson (1999), "Diffusion Indexes,"mimeo, ECARES, Universite Libre de Bruxelles. NBER Working Paper no. 6702.

Watson, M. W. and R. F. Engle (1983), "Alternative Algorithms for the Estimation of Dynamic Factors, MIMIC, and Varying Coefficient Regression Models," *Journal of Econometrics*, 23, 385–400.

CHAPTER 3

Macroeconomic Forecasting Using Many Predictors

Mark W. Watson

1. INTRODUCTION

The past twenty-five years have seen enormous intellectual effort and progress on the development of small-scale macroeconometric models. Indeed, standing in the year 2000, it is not too much of an overstatement to say that the statistical analysis of small macroeconometric models in a stationary environment is largely a completed research topic. In particular, we have complete theories of estimation, inference, and identification in stationary vector autoregressions (VARs). We have accumulated a vast amount of experience using these models for empirical analysis. Identified VARs have become the workhorse models for estimating the dynamic effects of policy changes and for answering questions about the sources of business cycle variability. Both univariate autoregressions and VARs are now standard benchmarks used to evaluate economic forecasts. Although work remains to be done, great progress has been made on the complications associated with nonstationarity, both in the form of the extreme persistence often found in macroeconomic time series and in detecting and modeling instability in economic relations. Threshold autoregressions and Markov switching models successfully capture much of the nonlinearity in macroeconomic relations, at least for countries such as the United States.[1]

Despite this enormous progress, it is also not too much of an overstatement to say that these small-scale macroeconometric models have had little effect on practical macroeconomic forecasting and policymaking.[2] There are several reasons for this, but the most obvious is the inherent defect of small models: they include only a small number of variables. Practical forecasters and policymakers find it useful to extract information from many more series than are typically included in a VAR.

[1] I will ignore small-scale "calibrated" dynamic models throughout this paper.

[2] Some might take issue with this broad assertion, and point to, for example, the VAR forecasting framework used for many years at the Federal Reserve Bank of Minneapolis. I would argue, however, that such examples are the exception rather than the rule.

This mismatch between standard macroeconometric models and real-world practice has led to two unfortunate consequences. First, forecasters have had to rely on informal methods to distill information from the available data, and their published forecasts reflect considerable judgment in place of formal statistical analysis. Forecasts are impossible to reproduce, and this makes economic forecasting a largely nonscientific activity. Because it is difficult to disentangle a forecaster's model and judgment, a predictive track record can tell us more about a person's insight than about the veracity of his or her model and the nature of the macroeconomy. The second unfortunate consequence is that formal small-scale macroeconometric models have little effect on day-to-day policy decisions, making these decisions more ad hoc and less predicable than if guided by the kind of empirical analysis that follows from careful statistical modeling.

In this paper I discuss a direction in macroeconometrics that explicitly incorporates information from a large number of macroeconomic variables into formal statistical models. The goal of this research is to use the wide range of economic variables that practical forecasters and macroeconomic policymakers have found useful, while at the same time maintaining the discipline and structure of formal econometric models. This research program is not new (several of the important contributions date from the 1970s and earlier), but it is still immature. There are few theoretical results and fewer empirical results, at least compared with small-scale models. Yet, the results that we do have suggest that there may be large payoffs from this "large-model" approach. The purpose of this paper is to summarize some of these results – particularly those relating to forecasting – to highlight some of the potential gains from these large models.

Section 2 begins the discussion by contrasting two approaches to macroeconometric forecasting. The first, the standard small-model method, constructs a forecasting equation with only a few variables; the second, a large-model method, uses information on a large number of variables. The main problem to be solved when constructing a small model is to choose the correct variables to include in the equation. This is the familiar problem of variable selection in regression analysis. Economic theory is of some help, but it usually suggests large categories of variables (money, interest rates, wages, stock prices, etc.), and the choice of a specific subset of variables then becomes a statistical problem. The large-model approach is again guided by economic theory for choosing categories of variables, and the statistical problem then becomes how to combine the information in this large collection of variables.

Sections 3 and 4 present empirical evidence on the relative merits of the small- and large-model approaches. Section 3 uses monthly U.S. data on 160 time series from 1959 to 1998 and examines the properties of regressions that include all of these 160 variables as regressors. The empirical analysis focuses on two questions. First, are the regressions characterized by a relatively few nonzero coefficients? If so, then only a few variables are needed, and there is no gain from looking across many series (except perhaps to choose the best variables to include in the model). Alternatively, if the regressions appear to

have a large number of small, but nonzero, coefficients, then a large-model approach that incorporates all of the variables may be more useful. Importantly, the empirical evidence summarized in Section 3 suggests that there are many nonzero regression coefficients, and thus provides support for the large-model approach. The second question taken up in this section focuses on the size of the marginal gains from including additional regressors. That is, are the predictors sufficiently correlated so that the bulk of their predictive content is contained in, say, only 30 variables, or are there large additional gains from including 50 or 100 or 150 variables? Here, the empirical evidence suggests that the marginal predictive gain of including additional regressors is a sharply decreasing function of the number of predictors, a result that is shown to be consistent with a factor analytic structure linking the regressors.

Section 4 uses these same data in a simulated out-of-sample forecasting experiment. This experiment focuses on twelve-month-ahead forecasts constructed over the 1970–1997 sample period. Forecasts are constructed from three small models and one large model. The small models are a univariate autoregression (which serves as a benchmark), a model that includes the variables present in a typical VAR (output, prices, interest rates, and commodity prices), and a model that includes a set of leading indicators suggested by previous forecasting comparisons. The large model is a version of the factor forecasting model developed in Stock and Watson (1998). This model summarizes the regressors using a few common factors that are used to construct the forecasts. For the majority of series studied, the factor model produces forecasts that are considerably more accurate than any of the small models.

The final section of the paper begins with an additional discussion of the results. It then continues with some speculative remarks about the future of this research program and concludes with a list of several outstanding problems.

2. SMALL AND LARGE MODELS

2.1. General Framework

This paper considers forecasting in a standard linear regression framework,

$$y_{t+1} = x_t'\beta + \varepsilon_{t+1}, \tag{2.1}$$

where y is the variable to be forecast using the vector of variables x, and the error term satisfies $E(\varepsilon_{t+1}|\{x_\tau, y_\tau\}_{\tau=-\infty}^{t}) = 0$. The dating of the variables in (2.1) emphasizes that it is a forecasting equation, and for notational ease the forecast horizon is set at one period. (The empirical work presented in what follows uses a twelve-month forecast horizon.) The equation does not include lagged values of y, and this too is for notational convenience. More substantively, (2.1) is specified as a time-invariant linear regression and so abstracts both from instability and from nonlinearity. The sample data are $\{y_t, x_t\}_{t=1}^{T}$, and the goal is to forecast y_{T+1}. The forecasts are constructed as $\hat{y}_{T+1} = x_T'\hat{\beta}_T$, where $\hat{\beta}_T$ is an estimate of β constructed from the sample information. Forecast

loss is quadratic, and so the forecast risk function is

$$R(\hat{\beta}_T, \beta) = \sigma_\varepsilon^2 + E[(\hat{\beta}_T - \beta)' x_T' x_T (\hat{\beta}_T - \beta)]. \tag{2.2}$$

Thus far, this is all standard.

The only nonstandard assumption concerns the number of regressors: x is an $n \times 1$ vector, where n is large. Thus, in contrast to standard large-T analysis of the regression model, here the analysis is carried out using a framework that assumes both large n and large T.

With this notation fixed, we now consider small-model and large-model approaches to the problem of estimating β and forecasting y.

2.2. Small-Model Approach

To interpret (2.1) as a small model, suppose that only k of the elements of the vector β are nonzero, where k is a small number. If the indices of the nonzero elements are known, then analysis is straightforward: x is partitioned as $x' = (x_1', x_2')$, where x_1 contains the k elements of x corresponding to the nonzero values of β and x_2 contains the remaining "irrelevant" variables. The regression coefficients are estimated by regressing y onto x_1, excluding x_2 from the regression. Imagining that k remains fixed as T grows large, this yields a consistent estimator of β, and the sampling error disappears from the limiting value of the forecasting risk: $\lim_{T\to\infty} R(\hat{\beta}_T, \beta) = \sigma_\varepsilon^2$.

The only remaining statistical problem becomes choosing the elements of x_1 from the vector x. This is the well-known variable-selection problem, and there are many standard methods for consistent variable selection. To highlight the key results for the forecasting problem, consider a special version of (2.1) with independent and identically distributed (iid.) $N(0, 1)$ disturbances and strictly exogenous orthogonal and standardized regressors (so that $T^{-1} \sum x_t' x_t = I_n$). Then, the ordinary least squares (OLS) estimators are $\hat{\beta}_{T,i} = T^{-1} \sum_{t=1}^{T} x_{it} y_t$, and the normality of the errors implies that $\hat{\beta}_{T,i} \sim \text{i.i.d.} N(\beta_i, T^{-1})$. Consider a simple variable-selection rule that chooses the r variables with the largest estimated coefficients. Suppose that r is fixed (not a function of T) with $r > k$. The key features of this variable-selection procedure follow from well-known results about the asymptotic behavior of order statistics. Let J denote the set of indices for the elements of β that are equal to zero. Let $b_{nT} = \max_{i \in J} |\hat{\beta}_{T,i}|$; then

$$\left(\frac{T}{2\log(n)}\right)^{1/2} b_{nT} \xrightarrow{as} 1$$

as n and $T \to \infty$ (Galambos, 1987). Thus, if $n = o(e^T)$, then $b_{nT} \xrightarrow{as} 0$.

Three important asymptotic properties follow directly from this result. First, with high probability the correct regressors (those in x_1) will be chosen for inclusion in the regression. Second, the estimated coefficients on the remaining r–k irrelevant selected variables will have coefficients that are close to zero. Finally, the implied forecasts have a limiting risk that is unaffected by sampling error in the estimated coefficients.

Taken together, this is good news for the small-model approach: if the number of variables that enter (2.1) is small and if the sample size is large, then sampling error associated with estimating the regression coefficients is likely to be small. Moreover, this conclusion continues to hold even if the variables chosen to enter (2.1) are determined by variable selection over a large number of possible models.

2.3. Large-Model Approaches

2.3.1. Factor Models

Macroeconomists naturally think of the comovement in economic time series as arising largely from a relatively few key economic factors such as productivity, monetary policy, and so forth. One way of representing this notion is in terms of a statistical factor model

$$x_t = \Lambda F_t + u_t, \tag{2.3}$$

which explains comovement among the variables x_t using a small number of latent factors F_t.[3] Although the classic factor model (Lawley and Maxwell, 1971) assumes that the elements of u_t are mutually uncorrelated both cross sectionally and temporally, this can be relaxed to allow temporal dependence among the us (Sargent and Sims, 1977; Geweke, 1977; and Engle and Watson, 1981), limited cross-sectional dependence (Chamberlain and Rothschild, 1983; and Connor and Korajczyk, 1986), or both (Forni et al., 1998; and Stock and Watson, 1998).

Pushing the factor model one step further, these latent factors might also explain the predictive relationship between x_t and y_{t+1}, so that

$$y_{t+1} = \alpha F_t + e_{t+1}, \tag{2.4}$$

where $E(e_{t+1}|\{y_\tau, x_\tau, F_\tau\}_{\tau=-\infty}^{t}) = 0$. As discussed in Stock and Watson (1998), (2.3) and (2.4) provide a potentially useful framework for large-model forecasting because they impose important constraints on the large number of regression coefficients in (2.1). Notably, if we write $E(F_t|x_t) = \gamma x_t$ (so that regression of F onto x is linear), then (2.4) implies that $E(y_{t+1}|x_t) = \alpha E(F_t|x_t) = \alpha\gamma x_t$, so that the regression coefficients β in (2.1) satisfy $\beta = \alpha\gamma$.

This model yields values of β that are quite different than in the small-model specification. In that model, most of the elements of β are zero, and only a few nonzero elements account for the predictive power in the regression. In the factor model, all of the elements of β are nonzero in general, but they are each small. This follows because the coefficients γ in the regression $E(F_t|x_t) = \gamma x_t$ will, in general, be $O(n^{-1})$ (the estimated value of F will be an average of the

[3] Versions of this framework have been explicitly used in empirical economics beginning with Sargent and Sims (1977) and Geweke (1977), but models such as this are implicit in earlier discussions of the business cycle. For example, Stock and Watson (1989) discuss the index of coincident indicators using a model much like (2.3).

n elements in x), and $\beta = \alpha\gamma$. Section 3 will examine sample versions of (2.1) for several macro variables and ask whether the distribution of the elements of β appears to be more consistent with the small model (many zero and a few large nonzero values) or the large model (many small nonzero values).

If the factor model can describe the data, then forecasting becomes a three-step process. First, an estimate of F_t, say $\hat{F}_t$, is constructed from the x data. Second, the coefficients α in (2.4) are estimated by regressing y_{t+1} onto $\hat{F}_t$. Finally, the forecast is formed as $\hat{y}_{T+1} = \hat{\alpha}\hat{F}_T$.

Somewhat remarkably, this procedure can work quite well, even when $\hat{F}_t$ is a very naïve estimator of F_t. For example, Stock and Watson (1998) show that, under limited cross-sectional and temporal dependence, that is, essentially $I(0)$ dependence both temporally and cross sectionally, with an identification condition on the factor loadings ($\Lambda'\Lambda$ must be well behaved), and when the number of regressors is sufficiently large [$n = O(T^\rho)$ for any $\rho > 0$], then $\hat{\alpha}\hat{F}_T - \alpha F_T \xrightarrow{p} 0$. That is, to first order, the feasible forecast constructed from the factor model is equal to the forecast that would be constructed if the true values of α and F_T were known.

An important practical problem is the determination of the number of factors to use in the analysis. There are two ways to do this. In the context of the forecasting problem, Stock and Watson (1998) propose a modification of the standard information criterion applied to the regression of y_{t+1} onto $\hat{F}_t$. They show that the resulting estimator is consistent and the resulting forecasts are asymptotically efficient. Alternatively, Bai and Ng (2001) propose estimators that are based on the fit of (2.3) and show consistency under assumptions similar to those used in Stock and Watson (1998).

2.3.2. Nonfactor Models

Large-model methods have also been proposed that do not rely on the factor framework. The most obvious is based on the OLS estimator of β, which is generally well defined when $n \leq T$. However, when the ratio of estimated parameters to observations (n/T) is large, the sampling error in the estimated coefficients will be large, and this suggests that significant improvements may be possible. From the classic result in Stein (1955), OLS estimates and forecasts are inadmissible when $n \geq 3$, and shrinkage estimators (such as the classic James–Stein estimator, James and Stein, 1960) dominate OLS. Although these results are largely irrelevant in the large-T and small-n model (as the risk of the OLS forecast converges to the risk of the infeasible known-β forecast), their relevance resurfaces in the large-T and large-n model. For example, when $n = \rho T$ with $0 < \rho < 1$, then the OLS estimator of β is not consistent and the sampling error in β continues to have a first-order effect on the forecast risk as $T \to \infty$. Shrinkage estimators can produce forecasts with asymptotic relative risks that dominate the OLS forecast, even asymptotically.

This idea is developed in Knox, Stock, and Watson (2001), where empirical Bayes estimators for β and forecasts from these estimators are constructed. They study parametric and nonparametric empirical Bayes estimators and provide

conditions under which the large-n assumption can be used to construct efficient forecasts (efficient in the sense that in large samples they achieve the same Bayes risk as the optimal forecasts constructed using knowledge of the distribution of the regression coefficients). To date, the empirical performance of these methods has not been systematically studied, and so it is premature to judge their usefulness for macroeconomic forecasting.

With this background we now proceed to a discussion of some empirical evidence on the relative merits of the small-model and large-model approaches to forecasting.

3. FULL-SAMPLE EMPIRICAL EVIDENCE

Two empirical questions are addressed in this section. First, in regressions such as (2.1), does it appear that only a few of the coefficients are nonzero? If so, then forecasts should be constructed using a small model. Alternatively, are there many nonzero (but perhaps small) regression coefficients? If so, then the large-model framework is appropriate. Second, and more generally, how does the regression's predictive R^2 change as the number of regressors is increased?

3.1. Data

These questions are investigated using data[4] on 160 monthly macroeconomic U.S. time series from 1959:1 through 1998:12. The data represent a wide range of macroeconomic activity and are usefully grouped into eight categories: Output and Real Income (twenty-one series); Employment and Unemployment (twenty-seven series); Consumption, Sales, and Housing (twenty-two series); Inventories and Orders (twenty-seven series); Prices and Wages (twenty-two series); Money and Credit (nine series); Interest Rates (nineteen series); and Exchange Rates and Stock Prices/Volume (thirteen series). These categories are perhaps overly broad, but it will be useful to have many series in each category, and so this coarse aggregation is necessary.

The data were transformed in three ways. First, many of the series are seasonally adjusted by the reporting agency. Second, the data were transformed to eliminate trends and obvious nonstationarities. For real variables, this typically involved transformation to growth rates (the first difference of logarithms), and for prices this involved transformation to changes in growth rates (the second difference of logarithms). Interest rates were transformed to first differences or to "spreads." Finally, some of the series contained a few large outliers associated with events such as labor disputes, other extreme events, or with data problems of various sorts. These outliers were identified as observations that differed from the sample median by more than six times the sample interquartile range, and these observations were dropped from the analysis.

A detailed description of the data can be found in Appendix A.

[4] These data were used in Stock and Watson (1999b).

3.2. Forecasting Regressions

Equation (2.1) was modified in two ways for the empirical analysis. First, the dependent variable was changed to focus on twelve-month-ahead forecasts, and second, autoregressive lags were added to the regression. The modified regression takes the form

$$y_{t+12}^{12} = x_t'\beta + \phi(L)y_t + e_t, \tag{3.1}$$

where y_t is the transformed variable of interest (as described in Section 3.1), and y_{t+12}^{12} denotes a further transformation of the variable appropriate for multistep forecasting. For example, if z_t denotes the raw (untransformed) value of the variable of interest and y_t is the monthly rate of growth of z_t, then $y_{t+12}^{12} = \ln(z_{t+12}/z_t)$ is the rate of growth over the forecast period. When y_t is the level of z_t, then $y_{t+12}^{12} = z_{t+12}$, and when y_t denotes the change in the monthly growth rate, then $y_{t+12}^{12} = \ln(z_{t+12}/z_t) - 12\ln(z_t/z_{t-1})$.

3.3. Empirical Distribution of t Statistics

Equation (3.1) was estimated by OLS for all 160 series in the data set. Each regression included a constant, current, and five lagged values of y_t, and an x vector containing the remaining 159 series (transformed as already described). Heteroskedastic and autoregressive consistent t statistics were computed for the estimated βs. This yielded 25,440 t statistics (160 regressions each, including 159 x regressors). Let τ_{ij} denote the t statistic for the ith β in the jth equation. Then, approximately, $\tau_{ij} \sim N(\mu_{ij}, 1)$ where μ_{ij} is the value of the noncentrality parameter for the t statistic. If $\mu_{ij} \neq 0$ then variable i is a useful predictor for y_{jt}; otherwise, variable i is not a useful predictor. We can determine the fraction of useful predictors by determining the fraction of nonzero values of μ_{ij}. Of course, μ_{ij} is not directly observed, but it is related to the t statistics by $\tau_{ij} = \mu_{ij} + \varepsilon_{ij}$, where (approximately) $\varepsilon_{ij} \sim N(0, 1)$ and is independent of μ_{ij}. This implies that the distribution of μ can then be estimated from the empirical distribution of τ by using deconvolution techniques. (The densities of τ and μ are related by $f_\tau(x) = \int \phi(x - s) f_\mu(s)ds$, where f_τ is the density of τ, f_μ is the density of μ, and ϕ is the standard normal density.)

Before looking at the results, suppose for a moment that the small-model assumptions described these data, so that a small number of β_is were nonzero. To be concrete, suppose that in each regression only six of the β_is were nonzero. In this case, over 96 percent (= 153/159) of the μs would equal zero, only 4 percent (= 6/159) would be nonzero, and these nonzero values would be large in absolute value. For example, if the regressors were orthonormal, then six values of $\mu_{ij} = 9$ would yield regressions with R^2s of approximately 50 percent. In contrast, if the large-model assumptions described the data, then most of the β_is would be nonzero, but small. This feature would be inherited by the noncentrality parameters μ_{ij}. For example, with othonormal regressors, values of μ_{ij} with $E(\mu_{ij}^2) = 1.7$ would yield regressions with R^2 values of 50 percent.

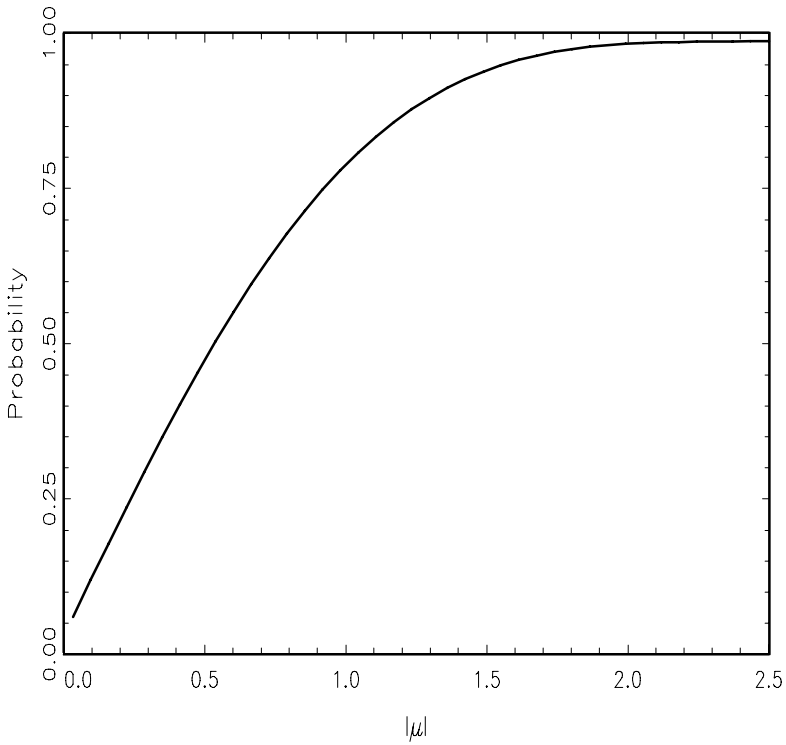


Figure 3.1. t statistic noncentrality parameter CDF (absolute value, $|\mu|$).

Figure 3.1 shows the estimated cumulative distribution function (CDF) of $|\mu_{ij}|$ estimated from the 25,440 values of τ_{ij}, using the deconvolution method outlined in Diggle and Hall (1993). The estimated CDF suggests that there are a large number of nonzero, but small, values of β. Over 55 percent of the noncentrality parameters exceed 0.5, 33 percent of the values lie between 0.5 and 1.0, 20 percent lie between 1.0 and 2.0, and only 2 percent of the values are above 2.0. These results suggest that the large-model assumptions are a better characterization of these data than the small-model assumptions.

Table 3.1 summarizes the distribution of the absolute values of the t statistics for selected variables. The first row of the table shows quartiles of the standard normal as a benchmark, and the next row shows the quartiles of the empirical distribution of all of the 25,440 t statistics from the estimated regressions. The median absolute t statistic was 0.84, which can be compared with a value of 0.67 for the normal distribution. Again, this suggests that a large fraction of the values of β in (3.1) are nonzero. The remaining rows of the table show the distribution across predictive regressions of the absolute t statistics for specific regressors. These distributions show whether the particular regressor has marginal predictive power for the range of macroeconomic variables in this data set. The first four entries are for the variables that are typically used

Table 3.1. *Distribution of absolute values of t statistics*

	Percentile		
	25th	50th	75th
Standard Normal	0.31	0.67	1.12
All Indicators	0.40	0.84	1.46
Indicator			
IP	0.66	1.16	1.65
PUNEW	0.28	0.63	1.06
FYFF	0.82	1.51	2.29
PMCP	0.45	1.02	1.82
LPHRM	0.70	1.22	1.93
LHU5	0.90	2.32	3.92
MOCMQ	0.32	0.66	1.12
PMDEL	0.30	0.68	1.19
MSONDQ	0.26	0.51	0.80
HSBR	0.63	1.27	2.10
FSPCOM	0.28	0.66	1.39
FM2DQ	0.54	1.12	1.79
SFYGT5	0.71	1.55	2.24
HHSNTN	0.69	1.44	2.36
PPSPR	0.64	1.39	2.21
LHUR	1.18	2.26	3.49
GMCNQ	0.40	0.75	1.31
FM1	0.34	0.80	1.33
FM2	0.30	0.64	1.11
PSM99Q	0.29	0.60	0.99
PWCMSA	0.26	0.54	0.90
MDU	0.50	0.97	1.46
IVMTQ	0.48	0.95	1.43

Note: The table shows the 25th, 50th, and 75th percentiles of the absolute values of the t statistics for the regressors shown in the first column across 160 forecasting regressions as described in Section 3.

in small VAR models: industrial production (IP), consumer price inflation (PUNEW), the federal funds rate (FYFF), and commodity prices (PMCP). With the exception of price inflation, these variables all perform better than the typical variable in the data set. The federal funds rate (FYFF) is particularly noteworthy, with a median absolute t statistic that exceeds 1.5. The next ten variables listed in the table correspond to the variables that make up the Conference Board's Index of Leading Indicators (previously published by the U.S. Department of Commerce).[5] Six of the ten variables are useful predictors for a large fraction of the variables considered. Notably, the number of newly unemployed (LHU5) had the largest median t statistic of all of the variables

[5] Because the data set used here did not include all of the Conference Board's indicators, there are two changes from the Conference Board's list. The number of new unemployment insurance claims is replaced by the number unemployed for less than five weeks (LHU5) and the five-year Treasury Bond rate is used in place of the ten-year rate in the term spread.

considered. This variable had an absolute t statistic that exceeded 1.0 in 72 percent of the regressions. In contrast, four of the Conference Board's indicators (the new orders and deliveries series, MOCMQ, PMDEL, and MSONDQ, and stock prices, FSPCOM) had little predictive power, and t-statistic distributions that are very close to that of the standard normal. The remaining rows of the table show results for other variables that have been widely discussed as leading indicators. The public-private interest rate spread (PPSPR) and the unemployment rate (LHUR) have large t statistics in a majority of the regressions. The nominal money supply (FM1 and FM2) is not very useful, although the real value of M2 (FM2DQ) was more useful. Unfilled orders and inventories have median t statistics that were nearly equal to unity, suggesting some useful predictive power in a large number of estimated regressions. Two other measures of commodity prices (PSM99Q and PWCMSA) are not useful predictors.

Of course, results based on these t statistics may be misleading for several reasons. What is probably most important is that they assume an $N(0,1)$ distribution for the sampling error in the t statistics, and this may be a poor approximation in the setting with a large number of regressors. However, these results are suggestive and are consistent with the other empirical results presented later that favor the large-model approach.

3.4. Prediction R^2s Using Different Numbers of Predictors

Although the empirical results in the last section suggest that "many" variables have marginal predictive content in (3.1), the results do not quantify the marginal gain from including, say, 50 variables instead of 25, or 150 instead of 100. If we let $R^2(k)$ denote the value of the population R^2 from the regression of y onto k elements of x, then the question is, How does $R^2(k)$ change as k increases? The answer is important for forecast design, but what is more important is that it provides information about the way macroeconomic variables interact. For example, suppose that each element of x contains important information about the future that cannot be gleaned from the other elements of x. (Housing starts in the Midwest contain some important information not contained in aggregate housing starts, the unemployment rate, or any of the other variables.) In this case $R^2(k)$ will be a steadily increasing function, with the amount of information increasing in proportion to the number of predictors.

Alternatively, suppose that macro variables interact in the simple low-dimensional way suggested by the factor model. In this case, each regressor will contain useful information about the values of the factors (and hence useful information about future values of y), but the marginal value of a particular regressor will depend on how many other variables have already been used. That is, if V_k denotes the variance of F_t conditional on the first k elements of x_t, then $V_k - V_{k-1}$ decreases as k increases. To see this most easily, consider the simplest case when there is only one factor, the uniquenesses, u in (2.3), are uncorrelated and homoskedastic, and the factor loadings, Λ in (2.3), are all unity. In this case the evolution of V_k is particularly simple: $V_k = [V_{k-1}/(\sigma_u^2 + V_{k-1})]V_{k-1}$, so that the variance decreases at a sharply decreasing rate as k increases. This implies

that $R^2(k)$ will increase at a sharply decreasing rate: there can be large gains from increasing the number of predictors from 5 to 25, but negligible gains from increasing the number of predictors from 100 to 120.

Stock and Watson (2001b) discuss the problem of estimating the population R^2 in equations such as (2.1) and (3.1) when the number of regressors is proportional to the sample size: $k = \rho T$, where $0 < \rho < 1$. In a classical version of (2.1), the usual degrees of freedom adjustment ($\bar{R}^2$) is all that is necessary to produce a consistent estimator of the population R^2. However, things are more complicated in a model such as (3.1) with predetermined but not exogenous regressors and an error term that is serially correlated and/or conditionally heteroskedastic. In this case an alternative estimator must be used. Here I use a split-sample estimator. Let $\hat{\beta}_1$ denote the least-squares estimator of β using the first half of the sample, and let $\hat{\beta}_2$ denote the estimator using the second half of the sample. Under general conditions, these two estimators will be approximately uncorrelated so that

$$n^{-1}\hat{\beta}_1'\hat{\beta}_2 \xrightarrow{p} \beta'\beta,\ T^{-1}\hat{\beta}_1' \sum_t x_t'x_t\hat{\beta}_2 \xrightarrow{p} E(\beta' x_t'x_t\beta)$$

and

$$R_S^2 = \left(\hat{\beta}_1 \sum x_t'x_t\hat{\beta}_2 / \sum y_t^2\right)$$

will provide a consistent estimator of the population R^2. Here, a partial R^2 version of this estimator is applied to (3.1) after the constant term and lagged values of y are controlled for.

The results are summarized in Figure 3.2. This figure plots the average estimated R^2 as a function of k.These values were computed as follows. First, a random set of k regressors was selected from the 160 available regressors. Equation (3.1) was then estimated for each of the y variables, and the split-sample R^2 (as just described) was computed after the lagged values of y were controlled for in the regression. This process was repeated 200 times, and the plot shows the average estimated R^2 as a function of k, where the average is computed across replications and across dependent variables.

The figure shows a sharp increase in R^2 as the number of regressors increases from $k = 5$ to $k = 50$ ($R^2(5) = .10$, $R^2(50) = .32$), but a much smaller increase as k increases from 100 to 150 ($R^2(100) = .35$, $R^2(150) = .37$). Thus, although many more variables appear to be useful than are typically used in small-scale models, the predictive component of the regressors is apparently common to many series in a way suggested by the factor model.

4. SIMULATED OUT-OF-SAMPLE EMPIRICAL EVIDENCE

This section uses a simulated real-time forecasting experiment to compare the forecasting performance of several small- and large-model forecasting methods. The experiment is similar to the experiments reported in Stock and Watson

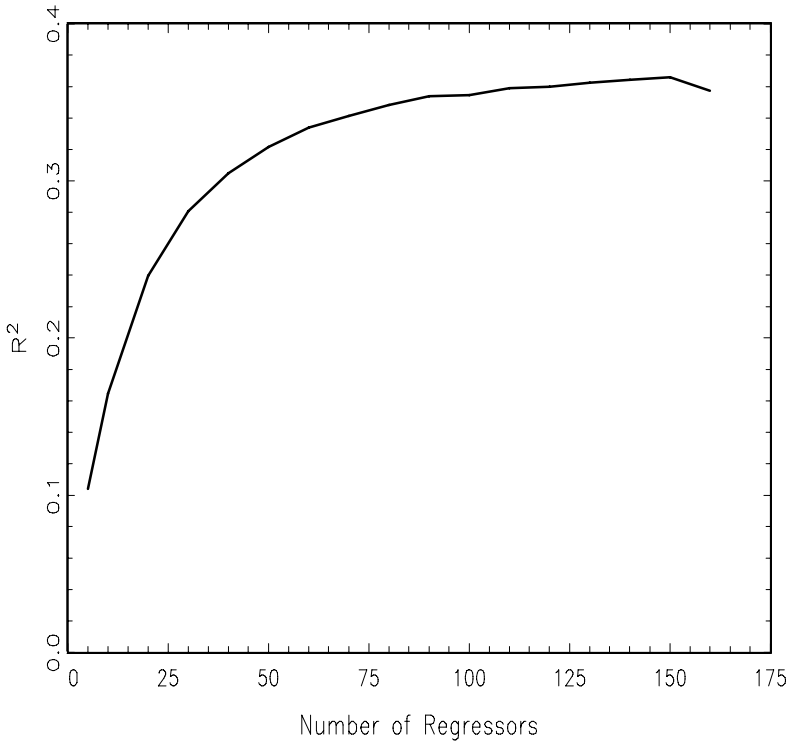


Figure 3.2. Predictive R^2 estimated average value.

(1999a, 2002), which studied forecasts of four measures of real activity and four measures of price inflation. Here forecasts for all 160 series in the data set are studied.

4.1. Experimental Design

The out-of-sample forecasting performance is simulated by recursively applying the forecasting procedures to construct twelve-month-ahead forecasts for each month of the sample beginning in $T = 1970{:}1$ through $T = 1997{:}12$. Small-model forecasts were computed using regression models of the form

$$y_{t+12}^{12} = \alpha + \beta(L)w_t + \phi(L)y_t + \varepsilon_{t+12}, \tag{4.1}$$

where y_{t+12}^{12} and y_t were defined in Section 3, and w_t is a (small) vector of predictors. Two versions of w_t were used. In the first, w_t includes the four variables typically included in monthly VARs (industrial production (IP), CPI inflation (PUNEW), the federal funds interest rate (FYFF), and commodity prices (PMCP)). In the second, w_t includes a set of eleven leading indicators that previous researchers have identified as useful predictors for real activity (three labor force indicators, LPHRM, LHEL, and LHNAPS; vendor

performance, PMDEL; capacity utilization, IPXMCA; housing starts, HSBR; the public-private interest and long-short interest rate spreads, PPSPR and TBSPR; long-term interest rates, FYGT10; the help-wanted index, exchange rates, EXRATE; and unfilled orders, MDU82). These are the leading indicators used in the Stock–Watson (1989) XLI and nonfinancial XLI. The forecasts from the first model are referred to as the VAR forecasts, even though they are computed directly from (4.1) instead of iterating the one-step VAR forward, and the second set of forecasts are called leading indicator forecasts. Forecasts from a univariate autoregressive model were also computed from (4.1) after w_t was eliminated.

In all of the models, the order of the lag polynomials $\beta(L)$ and $\phi(L)$ were determined recursively by Bayes Information Criterion (BIC). In the VAR, $\beta(L) = \sum_{i=0}^{q} \beta_i L^i$, where $0 \le q \le 5$ and $\beta(L) = 0$ was allowed; in the leading indicator model, $0 \le q \le 4$; in all of the models, $\phi(L) = \sum_{i=0}^{p} \phi_i L^i$, where $0 \le p \le 5$, and $\phi(L) = 0$ was allowed.

To be specific about the recursive forecasting experiment, consider the forecast dated T =1970:1. To construct this forecast, Equation (4.1) was estimated using data from t =1960:1 through t =1969:1, where the terminal date allows $y^{12}_{1970:1}$to be used as the dependent variable in the last period, and values of data before 1960:1 are used as lags for the initial periods. The lag lengths were determined by BIC using these sample data, and the regression coefficients were estimated conditional on these lag lengths. The forecast for $y^{12}_{1971:1}$was then constructed as $\hat{y}^{12}_{1971:1} = \hat{\alpha} + \hat{\beta}(L)w_{1970:1} + \hat{\phi}(L)y_{1970:1}$. To construct the forecast in 1970:2, the process was repeated with data from 1970:2 added to the data set. This experiment differs from what could have been computed in real time only because of revisions made in the historical data. (The data used here are from a 1999 release of the DRI Basic Macroeconomic Database.)

The experimental design for the factor model forecasts is similar. These forecasts were constructed from (4.1), with $\hat{F}_t$ replacing w_t. The estimated factors, $\hat{F}_t$, were computed recursively (so that the factors for the forecast constructed in 1970:1 were estimated using data on x_t from 1959:3 to 1970:1, where the initial two observations were lost because of second differencing of some of the variables). The factors were estimated by principal components, with missing values handled by a least-squares expectation-maximization algorithm as described in Stock and Watson (2002). Two sets of forecasts were constructed: The first used only contemporaneous values of the factors in the forecasting equation with the number of factors selected recursively by BIC; in the second, lags of the factors were allowed to enter the equation (and BIC was used to determine the number of factors and the number of lags). Because lags of the factors were nearly never chosen by BIC, the forecasts that allowed for lags were nearly identical to the forecasts that did not include lags. Thus, only results for the no-lag model are reported here.

The forecasting methods will be evaluated using the sample mean square error (MSE) over the simulated out-of-sample period: $\text{MSE}_{ij} = (336)^{-1} \sum_{T=1971:1}^{1998:12} (\hat{y}^{12}_{T,ij} - y^{12}_{T,i})^2$, where i indexes the series forecast, and j indexes the forecast method. Because the series differ in persistence and are

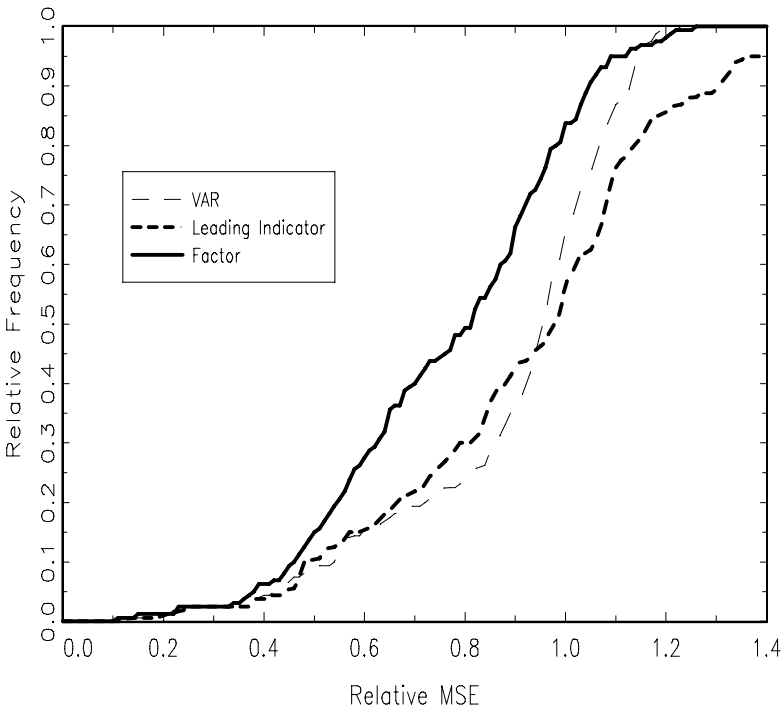


Figure 3.3. Relative MSE cumulative frequency distribution.

measured in different units, the MSEs relative to the MSE from the univariate autoregression will be reported. These will be referred to as relative MSEs.

4.2. Forecasting Performance of the Models

The forecasting performance of the models is summarized in Figure 3.3. This figure shows the cumulative relative frequency distribution of the relative MSEs for the VAR, leading indicator, and factor forecasting models. Each distribution summarizes the relative MSEs over the 160 variables that were forecast by each method. The figure shows that the factor forecasts clearly dominate the small-model forecasts. The median relative MSE for the factor model is 0.81 (indicating a 19 percent MSE gain relative to the univariate autoregression) compared with median values of 0.96 for the VAR and 0.98 for the leading indicator model. For 84 percent of the series, the factor model improved on the univariate autoregression, compared with 66 percent for the VAR and 57 percent for the leading indicator forecasts. For 44 percent of the series, the factor model produced more than a 25 percent MSE improvement relative to the univariate autoregression (so that the relative MSE was less than 0.75); only 22 percent of the VAR forecasts and 26 percent of the leading indicator forecasts achieved improvements this large.

Figure 3.4 presents a summary of the forecasting performance for the different categories of series. For each of the forecasting methods, the figure presents

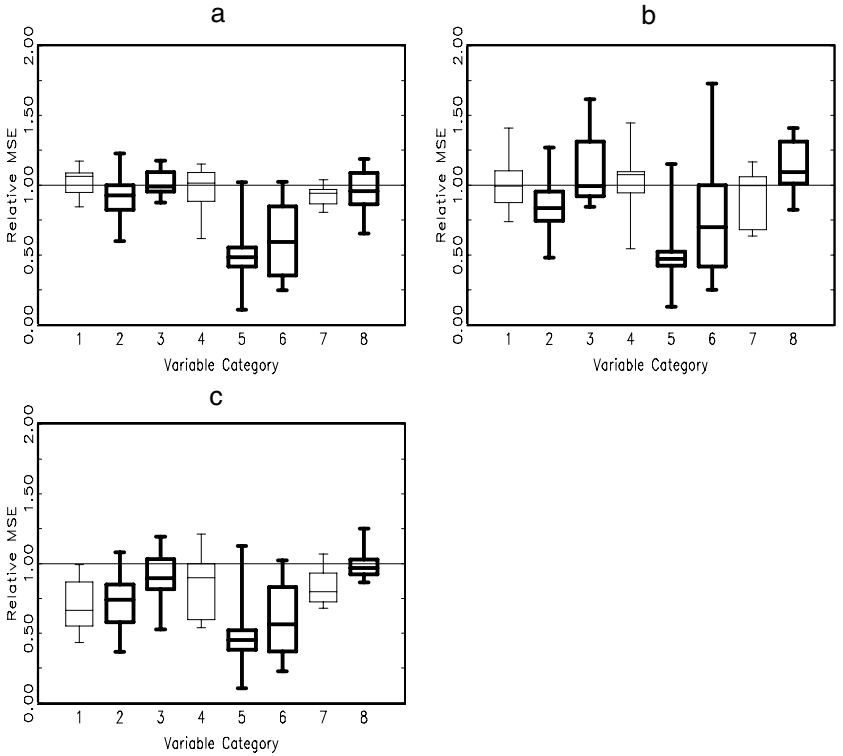


Figure 3.4. Relative MSE by forecast category: (a) VAR, (b) leading indicator, and (c) factor forecasts. Variable categories are as follows: 1. Output and Income; 2. Employment and Hours; 3. Consumption, Manufacturing, and Retail Saling and Housing; 4. Inventories and Inventory: Sales Ratios; 5. Prices and Wages; 6. Money and Credit; 7. Interest Rates; and 8. Exchange Rates, Stock Prices, and Volume.

box plots of the relative MSEs for each series category. (The outlines of the boxes are the 25th and 75th percentiles of the distribution, the median is the horizontal line in the box, and the vertical lines show the range of the distribution.) Looking first at the results for the factor model, Figure 3.4(c), we see that there are important differences in the forecastability of the series across categories. For example (and not surprisingly), there are negligible forecasting gains for series in the last category, which contains Exchange Rates and Stock Prices. In contrast, there are large gains for many of the other categories, notably the real activity variables in the Production and Employment categories, and the nominal variables making up the Prices, Wages, Money, and Credit categories. In all series categories, the median relative MSE is less than 1.0, and in five of the eight categories, the 75th percentile is less than 1.0.

The forecasting performance of the VAR and leading indicator models is more mixed. For the real variables in the first four categories, these models

Table 3.2. *Median of relative MSEs*

		Number of Factors		
Series Category	BIC Choice	1	2	3
Overall	0.81	0.88	0.77	0.81
Output and income	0.67	0.95	0.73	0.72
Employment and hours	0.75	0.78	0.69	0.74
Consumption and sales	0.90	0.98	0.85	0.88
Inventories and orders	0.90	0.92	0.81	0.86
Prices and wages	0.46	0.45	0.45	0.48
Money and credit	0.63	0.63	0.63	0.64
Interest rates	0.81	0.85	0.83	0.82
Exch. rates and stock prices	0.97	0.96	0.96	0.94

Note: This table shows the median of the MSE of the factor models for the series in the category listed in the first column. The MSEs are relative to the MSE for the univariate autoregression.

perform much worse than the factor model, and on average they do not improve on the univariate autoregression (the relative MSE across the variables in these categories is 1.0 for the VAR and 0.99 for the leading indicator model). In contrast, these models do yield improvements for wages, price, money, and credit that are roughly equal to the improvements produced by the factor forecasts.

4.3. Additional Discussion of the Factor Model

What accounts for the strong performance of the factor model and the uneven performance of the other models? Although I cannot provide a complete answer to this question, I can offer a few useful clues. Table 3.2 provides the first clue. It shows the relative MSE of several factor models: the model with the number of factors determined by BIC reported in Figures 3.3 and 3.4, and models with one, two, and three factors. Evidently, only a few factors are responsible for the model's forecasting performance. Two or three factors are useful for some categories of series, but only a single factor is responsible for the predictability of wages, prices, money, and credit.

The second clue is provided by Figure 3.5. It plots the estimated first factor along with the index of capacity utilization, one the variables used in the leading indicator model. These series are remarkably similar over much of the sample period. Major peaks and troughs of the estimated factor correspond closely to NBER business cycle dates. Apparently, the first factor is an index of real economic activity, and the forecasting results say that this real activity index is an important predictor of future price and wage inflation. Of course, this is just the well-known Phillips relation. Because both of the small models include good measures of the state of the real economy (industrial production in the VAR; capacity utilization, housing starts, and several of the other variables in the leading

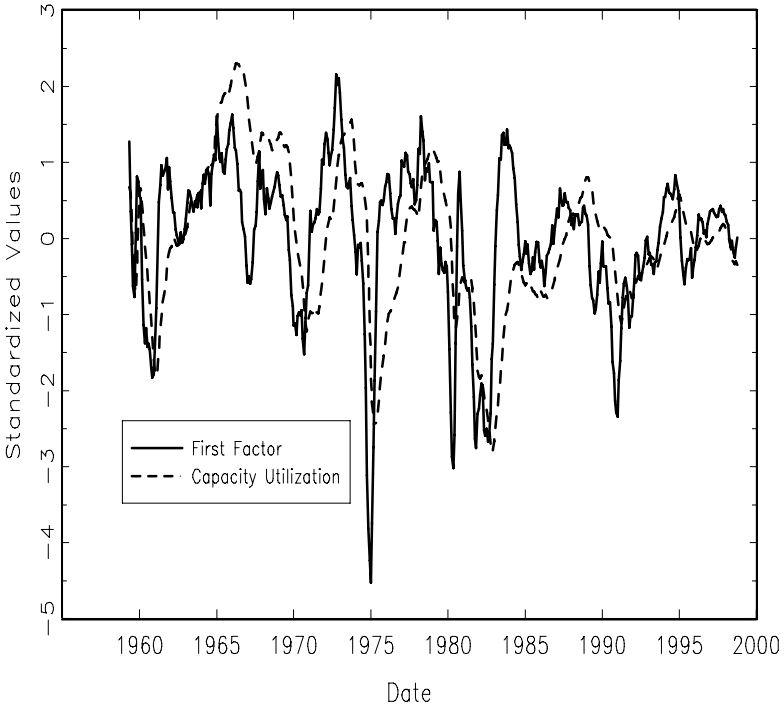


Figure 3.5. Factor 1 and index of capacity utilization.

indicator model), these models can exploit the Phillips relation to forecast future wage and price inflation, and this accounts for their good forecasting.

A close inspection of Figure 3.5 suggests that even though forecasts constructed from capacity utilization will be broadly similar to forecasts using the first factor, there may be important differences (compare, e.g., the series over 1975–1978). This raises the following question: When forecasting future inflation, do we need a large model, or should we forecast inflation using simple measures of real activity such as the unemployment rate, industrial production, or capacity utilization? The empirical evidence in Stock and Watson (1999a) provides a clear answer to this question, at least for typical measures of aggregate price inflation in the United States. An index of real activity constructed from a large number of variables performs better than any single series representing real activity. Apparently, there is enough idiosyncratic variation in standard activity measures such as the unemployment rate and capacity utilization that averaging these series to produce an activity index provides a clearer picture of aggregate activity and more accurate forecasts of inflation.

Although the first factor is easy to interpret, a complete understanding of the forecasting results requires an understanding of the second and third factors as well. Unfortunately, their interpretation and role in explaining future changes in real variables are an open question.

5. DISCUSSION

These empirical results raise several issues for economic forecasting and for macroeconometrics more generally. I use this section to highlight a few of these issues and to discuss a few of the large number of open research questions that they suggest.

Evaluations of the accuracy of macroeconomic forecasts (e.g., Zarnowitz and Braun, 1993) consistently find that "consensus" forecasts – averages of forecasts from many sources – are more accurate that individual forecasts. Averaging is a simple, but apparently very effective, large-model forecasting approach. How do the factor forecasts reported here compare with the consensus forecast benchmark? Differences in actual out-of-sample forecasting (which use real-time data) and the simulated out-of-sample forecasting carried out here (which use revised data) make a clean comparison difficult. However, a few calculations are suggestive. LaForte (2000) reports MSEs for the consensus forecast from the Survey of Professional Forecasters maintained by the Philadelphia Federal Reserve Bank (Croushore, 1993), and computes relative MSEs using univariate autoregressions recursively estimated using the real-time data set constructed by Croushore and Stark (1999). Over the sample period from 1969 to 1998, he reports relative MSEs of roughly 0.40 for aggregate price inflation (measured by the GNP/GDP price deflator) and the unemployment rate. (The precise value of the relative MSE depends on particular assumptions about the dates that forecasts were constructed and the specification of the univariate autoregression.) This value of 0.40 is only slightly larger than values for price inflation and the unemployment rate that were found here for the simulated forecasts using the factor model. This crude comparison suggests that the information aggregation in the factor model is roughly comparable with current best practice of using consensus forecasts.

Although these results are promising, a long list of open questions remains. Some are empirical and some are technical. Let me begin by listing two of the most obvious empirical questions. First, do the results reported here for the United States hold for other countries as well? Data limitations will make this question difficult to answer. For example, Marcellino, Stock, and Watson (2002) study forecasts of the unemployment rate, inflation, and short-term interest rates for European countries using data on over 500 series from 1982 to 1998. They find that estimated factors are highly significant for in-sample regressions, but they find inconclusive out-of-sample forecast rankings because of the short sample period. The second question concerns the stability of the forecasting relations. This is important in the United States, where the late 1990s seemed much different from the late 1970s, but is arguably more important for Europe, which has experienced enormous changes in the past decade.

The list of open technical questions is long. The problems of efficient estimation and inference that have been solved in small models remain open questions in large models. For example, the estimated factors used in the forecasting exercise reported here were constructed by the simplest of methods – principal

components. Although this estimator is consistent, undoubtedly more efficient estimators can be constructed. The empirical results in Table 3.2 indicated some substantial improvement from using models with a fixed number of factors rather than BIC selected factors, and this suggests that model selection procedures can be improved.

The existing theoretical results cover $I(0)$ models, but say nothing about integrated, cointegrated, and cotrending variables. We know that common long-run factors are important for describing macroeconomic data, and theory should be developed to handle these features in a large-model framework.

The difficult but important issues of nonlinearity and instability must also be addressed. Although some may scoff at the notion that there are large gains from modeling nonlinearity in macroeconomic time series, much of this (well-founded) skepticism comes from experience with small models. However, large-model results can be quite different. For example, in an experiment involving twelve-month-ahead forecasts for 215 macroeconomic time series, Stock and Watson (1999b) find that univariate autoregressions generally outperform standard nonlinear models (threshold autoregressions and artificial neural networks). Yet, in what can be interpreted as a large-model forecasting exercise, when forecasts from fifty nonlinear models were averaged, they outperformed any of the linear models, including combined models.

Temporal instability has already been mentioned as an open question in the context of the empirical work, but there are important technical questions as well. For example, are large-model methods more robust to instability than small-model methods? That is, does the cross-sectional averaging in large-model methods mitigate the effects of instability, and if so, what kinds of instability? There are already some results that relate to this question: Stock and Watson (1998) show that principal components estimators of factors remain consistent in the presence of some time variation in the factor loadings, but more general results are certainly possible and necessary.

Although this paper has focused on the problem of macroeconomic forecasting, the empirical results have more general implications for macroeconometric models. One need only consider the role that expectations play in theoretical models to appreciate this. There is an unfortunate bifurcation in the care in which expectations are handled in macroeconometric models. When expectations are explicitly incorporated in the models, and when interest focuses on a few key parameters, then empirical researchers are careful to use methods (such as instrumental variables) that are robust to the limited amount of information contained in small models. However, when expectations are implicitly included, as they are in most identified VARs, then researchers are much more cavalier, raising the possibility of large omitted variable biases. As Sims (1992) and Leeper, Sims, and Zha (1996) argue, this can have disastrous effects on inference and on policy advice. For example, Leeper et al. show that estimated effects of monetary policy on the macroeconomy decline sharply as they include more variables in their identified VAR to account for the central bank's forward-looking behavior. The largest model that they consider includes 18 variables, which is very large by conventional standards, but the results reported in

Section 3 suggest that there may be substantial increases in forecastability as the number of variables increases from, say, 18 to 50 or to 100. Thus there may still be large biases even in a VAR as large as the one constructed in their analysis. The construction of large-scale VARs or VAR-like models would seem to be a high priority for the large-model research program.

This paper has really been little more than a tease. It has pointed out important practical problems in the small-scale macroeconometric models that have been developed by researchers over the past twenty-five years. It has suggested that large models may solve many of these problems, so that formal statistical models can play a major role in economic forecasting and macroeconomic policy. A few theoretical results concerning large models were outlined. A set of empirical results were presented that suggest that these new models yield substantial improvements on small-scale models, and indeed may perform as well as the current best practice of using consensus forecasts. My hope is that others will find this tease intriguing enough to work on these problems and provide answers for the technical questions and empirical experience with the resulting new methods.

ACKNOWLEDGMENTS

This paper is based on research carried out jointly with James H. Stock, whom I thank for comments. Thanks also to Frank Diebold, Lars Hansen, and Lucrezia Reichlin for comments and Jean-Philippe LaForte for research assistance. This research was supported by the National Science Foundation (SBR-9730489).

APPENDIX A

Time series for sections 4 and 5

		Output and Income (Out)
1. IP	5	INDUSTRIAL PRODUCTION: TOTAL INDEX (1992=100, SA)
2. IPP	5	INDUSTRIAL PRODUCTION: PRODUCTS, TOTAL (1992=100, SA)
3. IPF	5	INDUSTRIAL PRODUCTION: FINAL PRODUCTS (1992=100, SA)
4. IPC	5	INDUSTRIAL PRODUCTION: CONSUMER GOODS (1992=100, SA)
5. IPCD	5	INDUSTRIAL PRODUCTION: DURABLE CONSUMER GOODS (1992=100, SA)
6. IPCN	5	INDUSTRIAL PRODUCTION: NONDURABLE CONSUMER GOODS (1992=100, SA)
7. IPE	5	INDUSTRIAL PRODUCTION: BUSINESS EQUIPMENT (1992=100, SA)
8. IPI	5	INDUSTRIAL PRODUCTION: INTERMEDIATE PRODUCTS (1992=100, SA)
9. IPM	5	INDUSTRIAL PRODUCTION: MATERIALS (1992=100, SA)
10. IPMD	5	INDUSTRIAL PRODUCTION: DURABLE GOODS MATERIALS (1992=100, SA)

11. IPMND	5	INDUSTRIAL PRODUCTION: NONDURABLE GOODS MATERIALS (1992=100, SA)
12. IPMFG	5	INDUSTRIAL PRODUCTION: MANUFACTURING (1992=100, SA)
13. IPD	5	INDUSTRIAL PRODUCTION: DURABLE MANUFACTURING (1992=100, SA)
14. IPN	5	INDUSTRIAL PRODUCTION: NONDURABLE MANUFACTURING (1992=100, SA)
15. IPMIN	5	INDUSTRIAL PRODUCTION: MINING (1992=100, SA)
16. IPUT	5	INDUSTRIAL PRODUCTION: UTILITIES (1992-=100, SA)
17. IPXMCA	1	CAPACITY UTIL RATE: MANUFACTURING, TOTAL (% OF CAPACITY, SA) (FRB)
18. PMI	1	PURCHASING MANAGERS' INDEX (SA)
19. PMP	1	NAPM PRODUCTION INDEX (PERCENT)
20. GMPYQ	5	PERSONAL INCOME (CHAINED) (SERIES #52) (BIL 92$, SAAR)
21. GMYXPQ	5	PERSONAL INCOME LESS TRANSFER PAYMENTS (CHAINED) (#51) (BIL 92$, SAAR)
		Employment and Hours (Emp)
22. LHEL	5	INDEX OF HELP-WANTED ADVERTISING IN NEWSPAPERS (1967=100;SA)
23. LHELX	4	EMPLOYMENT: RATIO; HELP-WANTED ADS: NO. UNEMPLOYED
24. LHEM	5	CIVILIAN LABOR FORCE: EMPLOYED, TOTAL (THOUS., SA)
25. LHNAG	5	CIVILIAN LABOR FORCE: EMPLOYED, NONAGRIC INDUSTRIES (THOUS., SA)
26. LHUR	1	UNEMPLOYMENT RATE: ALL WORKERS, 16 YEARS & OVER (%, SA)
27. LHU680	1	UNEMPLOY. BY DURATION: AVERAGE (MEAN) DURATION IN WEEKS (SA)
28. LHU5	1	UNEMPLOY. BY DURATION: PERSONS UNEMPL. LESS THAN 5 WKS (THOUS., SA)
29. LHU14	1	UNEMPLOY. BY DURATION: PERSONS UNEMPL. 5 TO 14 WKS (THOUS., SA)
30. LHU15	1	UNEMPLOY. BY DURATION: PERSONS UNEMPL. 15 WKS + (THOUS., SA)
31. LHU26	1	UNEMPLOY. BY DURATION: PERSONS UNEMPL. 15 TO 26 WKS (THOUS., SA)
32. LPNAG	5	EMPLOYEES ON NONAG. PAYROLLS: TOTAL (THOUS., SA)
33. LP	5	EMPLOYEES ON NONAG PAYROLLS: TOTAL, PRIVATE (THOUS., SA)
34. LPGD	5	EMPLOYEES ON NONAG. PAYROLLS: GOODS-PRODUCING (THOUS., SA)
35. LPMI	5	EMPLOYEES ON NONAG. PAYROLLS: MINING (THOUS., SA)
36. LPCC	5	EMPLOYEES ON NONAG. PAYROLLS: CONTRACT CONSTRUCTION (THOUS., SA)
37. LPEM	5	EMPLOYEES ON NONAG. PAYROLLS: MANUFACTURING (THOUS., SA)
38. LPED	5	EMPLOYEES ON NONAG. PAYROLLS: DURABLE GOODS (THOUS., SA)

39. LPEN	5	EMPLOYEES ON NONAG. PAYROLLS: NONDURABLE GOODS (THOUS., SA)
40. LPSP	5	EMPLOYEES ON NONAG. PAYROLLS: SERVICE-PRODUCING (THOUS., SA)
41. LPTU	5	EMPLOYEES ON NONAG. PAYROLLS: TRANS. & PUBLIC UTILITIES (THOUS., SA)
42. LPT	5	EMPLOYEES ON NONAG. PAYROLLS: WHOLESALE & RETAIL TRADE (THOUS., SA)
43. LPFR	5	EMPLOYEES ON NONAG. PAYROLLS: FINANCE, INSUR. & REAL ESTATE (THOUS., SA)
44. LPS	5	EMPLOYEES ON NONAG. PAYROLLS: SERVICES (THOUS., SA)
45. LPGOV	5	EMPLOYEES ON NONAG. PAYROLLS: GOVERNMENT (THOUS., SA)
46. LPHRM	1	AVG. WEEKLY HRS. OF PRODUCTION WKRS.: MANUFACTURING (SA)
47. LPMOSA	1	AVG. WEEKLY HRS. OF PROD. WKRS.: MFG., OVERTIME HRS. (SA)
48. PMEMP	1	NAPM EMPLOYMENT INDEX (PERCENT)
		Consumption, Manuf. and Retail Sales, and Housing (RTS)
49. HSFR	4	HOUSING STARTS: NONFARM(1947–1958); TOTAL FARM & NONFARM (THOUS., SA)
50. HSNE	4	HOUSING STARTS: NORTHEAST (THOUS.U.) SA
51. HSMW	4	HOUSING STARTS: MIDWEST(THOUS.U.) SA
52. HSSOU	4	HOUSING STARTS: SOUTH (THOUS.U.) SA
53. HSWST	4	HOUSING STARTS: WEST (THOUS.U.) SA
54. HSBR	4	HOUSING AUTHORIZED: TOTAL NEW PRIV HOUSING UNITS (THOUS., SAAR)
55. HMOB	4	MOBILE HOMES: MANUFACTURERS' SHIPMENTS (THOUS.OF UNITS, SAAR)
56. MSMTQ	5	MANUFACTURING & TRADE: TOTAL (MIL OF CHAINED 1992 DOLLARS) (SA)
57. MSMQ	5	MANUFACTURING & TRADE: MANUFACTURING; TOTAL (MIL OF CH. 1992 DOLLARS) (SA)
58. MSDQ	5	MANUFACTURING & TRADE: MFG; DURABLE GOODS (MIL OF CH. 1992 DOLLARS) (SA)
59. MSNQ	5	MANUFACT. & TRADE: MFG; NONDURABLE GOODS (MIL OF CHAINED 1992$) (SA)
60. WTQ	5	MERCHANT WHOLESALERS: TOTAL (MIL OF CHAINED 1992 DOLLARS) (SA)
61. WTDQ	5	MERCHANT WHOLESALERS: DURABLE GOODS TOTAL (MIL OF CH. 1992 DOLLARS) (SA)
62. WTNQ	5	MERCHANT WHOLESALERS: NONDURABLE GOODS (MIL OF CHAINED 1992$) (SA)
63. RTQ	5	RETAIL TRADE: TOTAL (MIL OF CHAINED 1992 DOLLARS) (SA)
64. RTNQ	5	RETAIL TRADE: NONDURABLE GOODS (MIL OF 1992 DOLLARS) (SA)
65. HHSNTN	1	U. OF MICH. INDEX OF CONSUMER EXPECTATIONS (BCD-83)
66. GMCQ	5	PERSONAL CONSUMPTION EXPEND (CHAINED)–TOTAL (BIL 92$, SAAR)

67. GMCDQ	5	PERSONAL CONSUMPTION EXPEND (CHAINED)–TOTAL DURABLES (BIL 92$, SAAR)
68. GMCNQ	5	PERSONAL CONSUMPTION EXPEND (CHAINED)–NONDURABLES (BIL 92$, SAAR)
69. GMCSQ	5	PERSONAL CONSUMPTION EXPEND (CHAINED)–SERVICES (BIL 92$, SAAR)
70. GMCANQ	5	PERSONAL CONS EXPEND (CHAINED)–NEW CARS (BIL 92$, SAAR)
		Real Inventories and Inventory: Sales Ratios (Inv)
71. IVMTQ	5	MANUFACTURING & TRADE INVENTORIES: TOTAL (MIL OF CHAINED 1992)(SA)
72. IVMFGQ	5	INVENTORIES, BUSINESS, MFG (MIL OF CHAINED 1992 DOLLARS, SA)
73. IVMFDQ	5	INVENTORIES, BUSINESS, DURABLES (MIL OF CHAINED 1992 DOLLARS, SA)
74. IVMFNQ	5	INVENTORIES, BUSINESS, NONDURABLES (MIL OF CHAINED 1992 DOLLARS, SA)
75. IVWRQ	5	MANUFACTURING & TRADE INV: MERCHANT WHOLESALERS (MIL OF CH. 1992 $) (SA)
76. IVRRQ	5	MANUFACTURING & TRADE INV: RETAIL TRADE (MIL OF CHAINED 1992 DOLLARS) (SA)
77. IVSRQ	2	RATIO FOR MFG & TRADE: INVENTORY:SALES (CHAINED 1992 DOLLARS, SA)
78. IVSRMQ	2	RATIO FOR MFG & TRADE: MFG; INVENTORY:SALES (87$) (SA)
79. IVSRWQ	2	RATIO FOR MFG & TRADE: WHOLESALER; INVENTORY:SALES (87$) (SA)
80. IVSRRQ	2	RATIO FOR MFG & TRADE:RETAIL TRADE;INVENTORY:SALES(87$) (SA)
81. PMNO	1	NAPM NEW ORDERS INDEX (PERCENT)
82. PMDEL	1	NAPM VENDOR DELIVERIES INDEX (PERCENT)
83. PMNV	1	NAPM INVENTORIES INDEX (PERCENT)
84. MOCMQ	5	NEW ORDERS (NET)-CONSUMER GOODS & MATERIALS, 1992 DOLLARS (BCI)
85. MDOQ	5	NEW ORDERS, DURABLE GOODS INDUSTRIES, 1992 DOLLARS (BCI)
86. MSONDQ	5	NEW ORDERS, NONDEFENSE CAPITAL GOODS, IN 1992 DOLLARS (BCI)
87. MO	5	MFG NEW ORDERS: ALL MANUFACTURING INDUSTRIES, TOTAL (MIL$, SA)
88. MOWU	5	MFG NEW ORDERS: MFG INDUSTRIES WITH UNFILLED ORDERS(MIL$, SA)
89. MDO	5	MFG NEW ORDERS: DURABLE GOODS INDUSTRIES, TOTAL (MIL$, SA)
90. MDUWU	5	MFG NEW ORDERS: DURABLE GOODS INDUSTRIES WITH UNFILLED ORDERS(MIL$, SA)
91. MNO	5	MFG NEW ORDERS: NONDURABLE GOODS INDUSTRIES, TOTAL (MIL$, SA)
92. MNOU	5	MFG NEW ORDERS: NONDURABLE GOODS INDUSTRIES WITH UNFILLED ORDERS(MIL$, SA)
93. MU	5	MFG UNFILLED ORDERS: ALL MANUFACTURING INDUSTRIES, TOTAL (MIL$, SA)

94. MDU	5	MFG UNFILLED ORDERS: DURABLE GOODS INDUSTRIES, TOTAL (MIL$, SA)
95. MNU	5	MFG UNFILLED ORDERS: NONDURABLE GOODS INDUSTRIES, TOTAL (MIL$, SA)
96. MPCON	5	CONTRACTS & ORDERS FOR PLANT & EQUIPMENT (BIL$, SA)
97. MPCONQ	5	CONTRACTS & ORDERS FOR PLANT & EQUIPMENT IN 1992 DOLLARS (BCI)
		Prices and Wages (PWG)
98. LEHCC	6	AVG HR EARNINGS OF CONSTR WKRS: CONSTRUCTION ($, SA)
99. LEHM	6	AVG HR EARNINGS OF PROD WKRS: MANUFACTURING ($, SA)
100. PMCP	1	NAPM COMMODITY PRICES INDEX (PERCENT)
101. PWFSA	6	PRODUCER PRICE INDEX: FINISHED GOODS (82=100, SA)
102. PWFCSA	6	PRODUCER PRICE INDEX: FINISHED CONSUMER GOODS (82=100, SA)
103. PWIMSA	6	PRODUCER PRICE INDEX: INTERMED MAT. SUPPLIES & COMPONENTS(82=100, SA)
104. PWCMSA	6	PRODUCER PRICE INDEX: CRUDE MATERIALS (82=100, SA)
105. PSM99Q	6	INDEX OF SENSITIVE MATERIALS PRICES (1990=100)(BCI-99A)
106. PUNEW	6	CPI-U: ALL ITEMS (82–84=100, SA)
107. PU83	6	CPI-U: APPAREL & UPKEEP (82–84=100, SA)
108. PU84	6	CPI-U: TRANSPORTATION (82–84=100, SA)
109. PU85	6	CPI-U: MEDICAL CARE (82–84=100, SA)
110. PUC	6	CPI-U: COMMODITIES (82–84=100, SA)
111. PUCD	6	CPI-U: DURABLES (82–84=100, SA)
112. PUS	6	CPI-U: SERVICES (82–84=100, SA)
113. PUXF	6	CPI-U: ALL ITEMS LESS FOOD (82–84=100, SA)
114. PUXHS	6	CPI-U: ALL ITEMS LESS SHELTER (82–84=100, SA)
115. PUXM	6	CPI-U: ALL ITEMS LESS MEDICAL CARE (82–84=100, SA)
116. GMDC	6	PCE, IMPL PR DEFL: PCE (1987=100)
117. GMDCD	6	PCE, IMPL PR DEFL: PCE; DURABLES (1987=100)
118. GMDCN	6	PCE, IMPL PR DEFL: PCE; NONDURABLES (1987=100)
119. GMDCS	6	PCE, IMPL PR DEFL: PCE; SERVICES (1987=100)
		Money and Credit Quantity Aggregates (Mon)
120. FM1	6	MONEY STOCK: M1(CURR, TRAV.CKS, DEM DEP, OTHER CK'ABLE DEP) (BIL$, SA)
121. FM2	6	MONEY STOCK: M2(M1+ON RPS, ER$, G/P & B/D MMMFS & SAV & SM TM DEP (B$, SA)
122. FM3	6	MONEY STOCK: M3(M2+LG TIME DEP,TERM RP'S & INST ONLY MMMFS) (BIL$, SA)
123. FM2DQ	5	MONEY SUPPLY-M2 IN 1992 DOLLARS (BCI)
124. FMFBA	6	MONETARY BASE, ADJ FOR RESERVE REQUIREMENT CHANGES (MIL$, SA)
125. FMRRA	6	DEPOSITORY INST RESERVES: TOTAL, ADJ FOR RESERVE REQ CHGS (MIL$, SA)
126. FMRNBC	6	DEPOSITORY INST RESERVES: NONBOR + EXT CR, ADJ RES REQ CGS (MIL$, SA)

127. FCLNQ	5	COMMERCIAL & INDUSTRIAL LOANS OUSTANDING IN 1992 DOLLARS (BCI)
128. FCLBMC	1	WKLY RP LG COM'L BANKS: NET CHANGE COM'L & INDUS LOANS(BIL$, SAAR)
		Interest Rates (Int)
129. FYFF	2	INTEREST RATE: FEDERAL FUNDS (EFFECTIVE) (% PER ANNUM, NSA)
130. FYCP90	2	INTEREST RATE: 90 DAY COMMERCIAL PAPER, (AC) (% PER ANN, NSA)
131. FYGM3	2	INTEREST RATE: U.S.TREASURY BILLS, SEC MKT, 3–MO. (% PER ANN, NSA)
132. FYGM6	2	INTEREST RATE: U.S.TREASURY BILLS, SEC MKT, 6–MO. (% PER ANN, NSA)
133. FYGT1	2	INTEREST RATE: U.S.TREASURY CONST MATURITIES, 1–YR. (% PER ANN, NSA)
134. FYGT5	2	INTEREST RATE: U.S.TREASURY CONST MATURITIES, 5–YR. (% PER ANN, NSA)
135. FYGT10	2	INTEREST RATE: U.S.TREASURY CONST MATURITIES, 10–YR. (% PER ANN, NSA)
136. FYAAAC	2	BOND YIELD: MOODY'S AAA CORPORATE (% PER ANNUM)
137. FYBAAC	2	BOND YIELD: MOODY'S BAA CORPORATE (% PER ANNUM)
138. FYFHA	2	SECONDARY MARKET YIELDS ON FHA MORTGAGES (% PER ANNUM)
139. SFYGM3	1	Spread FYGM3 – FYFF
140. SFYGM6	1	Spread FYGM6 – FYFF
141. SFYGT1	1	Spread FYGT1 – FYFF
142. SFYGT5	1	Spread FYGT5 – FYFF
143. SFYAAAC	1	Spread FYAAAC – FYFF
144. SFYBAAC	1	Spread FYBAAC – FYFF
145. SFYFHA	1	Spread FYFHA – FYFF
146. PPSPR	1	Public-Private Spread FYCP90–FYGM3
147. TBSPR	1	Term Spread FYGT10–FYGT1
		Exchange Rates, Stock Prices and Volume (ESP)
148. FSNCOM	5	NYSE COMMON STOCK PRICE INDEX: COMPOSITE (12/31/65=50)
149. FSPCOM	5	S&P'S COMMON STOCK PRICE INDEX: COMPOSITE (1941–43=10)
150. FSPIN	5	S&P'S COMMON STOCK PRICE INDEX: INDUSTRIALS (1941–43=10)
151. FSPCAP	5	S&P'S COMMON STOCK PRICE INDEX: CAPITAL GOODS (1941–43=10)
152. FSPUT	5	S&P'S COMMON STOCK PRICE INDEX: UTILITIES (1941–43=10)
153. FSDXP	1	S&P'S COMPOSITE COMMON STOCK: DIVIDEND YIELD (% PER ANNUM)
154. FSPXE	1	S&P'S COMPOSITE COMMON STOCK: PRICE-EARNINGS RATIO (%,NSA)
155. EXRUS	5	UNITED STATES;EFFECTIVE EXCHANGE RATE (MERM)(INDEX NO.)

156. EXRGER	5	FOREIGN EXCHANGE RATE: GERMANY (DEUTSCHE MARK PER U.S.$)
157. EXRSW	5	FOREIGN EXCHANGE RATE: SWITZERLAND (SWISS FRANC PER U.S.$)
158. EXRJAN	5	FOREIGN EXCHANGE RATE: JAPAN (YEN PER U.S.$)
159. EXRUK	5	FOREIGN EXCHANGE RATE: UNITED KINGDOM (CENTS PER POUND)
160. EXRCAN	5	FOREIGN EXCHANGE RATE: CANADA (CANADIAN $ PER U.S.$)

Note: This appendix lists the time series used in the empirical analysis in Sections 4 and 5. The series were either taken directly from the DRI/McGraw-Hill Basic Economics database, in which case the original mnemonics are used, or they were produced by authors calculations based on data from that database, in which case the authors' calculations and original DRI/McGraw-Hill series mnemonics are summarized in the data description field. Following the series name is a transformation code and a short data description. The value of the transformation code indicates the transformation discussed in subsection 3.1. The transformations are (1) level of the series; (2) first difference; (3) second difference; (4) logarithm of the series; (5) first difference of the logarithm; and (6) second difference of the logarithm. The following abbreviations appear in the data descriptions: SA = seasonally adjusted; NSA = not seasonally adjusted; SAAR = seasonally adjusted at an annual rate; FRB = Federal Reserve Board; AC = authors' calculations.

References

Bai, J. and S. Ng (2002), "Determining the Number of Factors in Approximate Factor Models," *Econometrica*, 70:1, 191–221.

Chamberlain, G. and M. Rothschild (1983), "Arbitrage Factor Structure, and Mean-Variance Analysis of Large Asset Markets," *Econometrica*, 51(5), 1281–1304.

Connor, G. and R. A. Korajczyk (1986), "Performance Measurement with the Arbitrage Pricing Theory," *Journal of Financial Economics*, 15, 373–394.

Croushore, D. (1993), "Introducing the Survey of Professional Forecasters," *Federal Reserve Bank of Philadelphia Business Review*, November/December, 3–13.

Croushore, D. and T. Stark (1999), "A Real-Time Data Set for Macroeconomists," Working Paper 99-4, Philadelphia Federal Reserve Bank.

Diggle, P. J. and P. Hall (1993), "A Fourier Approach to Nonparametric Deconvolution of a Density Estimate," *Journal of the Royal Statistical Society, Series B*, 55, 523–531.

Engle, R. F. and M. W. Watson (1981), "A One-Factor Multivariate Time Series Model of Metropolitan Wage Rates," *Journal of the American Statistical Association*, 76(376), 774–781.

Forni, M., M. Hallin, M. Lippi, and L. Reichlin (2000), "The Generalized Dynamic Factor Model: Identification and Estimation," *The Review of Economics and Statistics*, 82:4, 540–554.

Galambos, J. (1987), *The Asymptotic Theory of Extreme Order Statistics*, 2nd ed. Malabar, FL: Krieger.

Geweke, J. (1977), "The Dynamic Factor Analysis of Economic Time Series," in *Latent Variables in Socio-Economic Models*. (ed. by D. J. Aigner and A. S. Goldberger), Amsterdam: North-Holland, Chapter 19.

James, W. and C. Stein (1960), "Estimation with Quadratic Loss," *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, 1, 361–379.

Knox, T., J. H. Stock, and M. W. Watson (2001), "Empirical Bayes Forecasts of One Time Series Using Many Predictors," NBER Technical Working Paper 269.

LaForte, J. P (2000), "The Relative Accuracy of the Survey of Professional Forecasters," Research Memorandum, Princeton University.

Lawley, D. N. and A. E. Maxwell (1971), *Factor Analysis as a Statistical Method*. New York: American Elsevier.

Leeper, E. M., C. A. Sims, and T. Zha (1996), "What Does Monetary Policy Do?" *Brookings Papers on Economic Activity*, 2, 1–78.

Marcellino, M., J. H. Stock, and M. W. Watson (2002), "Macroeconomic Forecasting in the Euro Area: Country Specific Versus Area-Wide Information," *European Economic Review*, in press.

Sargent, T. J. and C. A. Sims (1977), "Business Cycle Modeling Without Pretending to Have Too Much a Priori Economic Theory," in *New Methods in Business Cycle Research*, (ed. by C. Sims et al.), Minneapolis: Federal Reserve Bank of Minneapolis.

Sims, C. A. (1992), "Interpreting the Macroeconomic Time Series Facts: The Effects of Monetary Policy," *European Economic Review*, 36, 975–1001.

Stein, C. (1955), "Inadmissibility of the Usual Estimator for the Mean of Multivariate Normal Distribution," *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, 1, 197–206.

Stock, J. H. and M. W. Watson (1989), "New Indexes of Coincident and Leading Economic Indicators," *NBER Macroeconomics Annual*, 351–393.

Stock, J. H. and M. W. Watson (1998), "Diffusion Indexes," NBER Working Paper W6702.

Stock, J. H. and M. W. Watson (1999a), "Forecasting Inflation," *Journal of Monetary Economics*, (44)2, 293–335.

Stock, J. H. and M. W. Watson (1999b), "A Comparison of Linear and Nonlinear Univariate Models for Forecasting Macroeconomic Time Series," in *Cointegration, Causality, and Forecasting: A Festschrift in Honour of Clive W. J. Granger* (ed. by R. F. Engle and H. White), Oxford: Oxford University Press.

Stock, J. H. and M. W. Watson (2001b), "The Limits of Predictability," in preparation.

Stock, J. H. and M. W. Watson (2002), "Macroeconomic Forecasting Using Diffusion Indexes," *Journal of Business and Economic Statistics*, in press.

Zarnowitz, V. and P. Braun (1993), "Twenty-two Years of the NBER-ASA Quarterly Economic Outlook Surveys: Aspects and Comparisons of Forecasting Performance," in *Business Cycles, Indicators, and Forecasting*, (ed. by J. H. Stock and M. W. Watson), NBER Studies in Business Cycles, Vol. 28, Chicago: University of Chicago Press.

"Big Data" Dynamic Factor Models for Macroeconomic Measurement and Forecasting

A Discussion of the Papers by Lucrezia Reichlin and by Mark W. Watson

Francis X. Diebold

1. BIG DATA

The Reichlin and Watson papers are just what we have come to expect from their authors: practical and pragmatic, yet grounded in rigorous theory. In short, good science.

Recently, much good science, whether physical, biological, or social, has been forced to confront – and has often benefited from – the "Big Data" phenomenon. Big Data refers to the explosion in the quantity (and sometimes, quality) of available and potentially relevant data, largely the result of recent and unprecedented advancements in data recording and storage technology. In this new and exciting world, sample sizes are no longer fruitfully measured in "number of observations," but rather in, say, megabytes. Even data accruing at the rate of several gigabytes per day are not uncommon. Economics examples include microeconometric analyses of consumer choice, which have been transformed by the availability of huge and detailed data sets collected by checkout scanners, and financial econometric analyses of asset return dynamics, which have been similarly transformed by the availability of tick-by-tick data for thousands of assets.

From the vantage point of the examples just sketched, the Reichlin and Watson papers do not analyze *really* Big Data, but they certainly represent a movement of macroeconometrics in that direction. In traditional small-data macroeconometric environments such as those explored by Stock and Watson (1989), one might work with a vector autoregression involving, say, four or five macroeconomic indicators measured over 150 quarters. In contrast, Reichlin and Watson work with roughly 500 indicators measured over 500 months. In the not-too-distant future, we will be working with thousands of indicators, with many measured at daily or higher frequencies.

A canonical problem in the analysis of Big Data is how to make tractable an "X matrix" (in regression parlance) of dimension $T \times K$, when both T and K are very large. In the applications at hand, the variables in X are used to extract or forecast the state of macroeconomic activity (f), on which they depend. The latent factor f may be the object of intrinsic interest (Reichlin),

or it may be used to forecast some other variable, say, $y = g(f)$ (Watson). Many approaches proceed by somehow reducing the number of columns of X. Variable-selection methods, for example, whether based on the traditional tools of inference such as t and F tests or on more recently developed criteria such as Akaike Information Criterion, amount to strategies for eliminating columns of X. The principal component methods used by Reichlin and Watson are rather more sophisticated, not requiring a sharp "in" or "out" decision for each variable, but rather allowing all variables to contribute to an extraction or forecast. Of course, replacing a large set of variables with a small set of their first few principal components cannot, in general, be done without substantial information loss. It is truly fortunate for macroeconomists and financial economists that our data are often well approximated by low-dimensional factor structures, in which case replacing large sets of variables by a few principal components is not only convenient but also legitimate (in the sense of little information loss).

2. FACTOR STRUCTURE AND REGIME SWITCHING

Let us elaborate on the idea of factor structure. To do so it will be helpful to recall what I call the linear tradition in macroeconomics and business cycle analysis, which emphasizes the modeling and interpretation of comovement among macroeconomic aggregates, as in Burns and Mitchell (1946). Part of the modern econometric distillation of that tradition is the vector autoregression, a linear model that captures comovement by allowing lags of variable i to affect variable j, for all i and j, and by allowing for contemporaneous correlation across shocks. However, the vector autoregression alone does not provide a viable description of large sets of macroeconomic indicators; degrees of freedom would soon be exhausted. Hence the appeal of dynamic factor structures (Sargent and Sims, 1977 and Geweke, 1977), reflecting the recognition that comovements among macroeconomic indicators likely arise from partial dependence on common shocks. In a one-factor model, which is the leading case in practice, there is just one common shock. Hence the behavior of each of a potentially large set of K variables is qualitatively similar to the behavior of just one variable, the common factor.

There is also, however, a distinct nonlinear tradition in macroeconomics and business cycle analysis, which emphasizes the idea of regime switching, namely that expansions and contractions may be usefully treated as different probabilistic objects, with turning points naturally defined as switching times. This view is clearly delineated in classics such as Burns and Mitchell (1946) and is embodied in modern regime-switching models, particularly the popular Markov-switching model of Hamilton (1989). In Hamilton's model, conditional densities are governed by a parameter vector whose value depends on the state (expansion or contraction, say), with state transitions governed by a first-order Markov process.

The linear and nonlinear traditions have matured largely in isolation, but they are in no way contradictory, and accurate business cycle measurement and forecasting may require elements of both. Hence, Diebold and Rudebusch (1996) propose a dynamic factor model in which the factor may display Markov switching. For concreteness, and because it will feature prominently in the sequel, I sketch a simple one-factor model with first-order autoregressive dynamics and a mean growth rate that switches across expansions and contractions. In an obvious notation,

$$x_t = \beta + \lambda f_t + u_t, \tag{2.1}$$

where x is a vector of covariance stationary indicators, β is a vector of constants, λ is a vector of factor loadings, f is a scalar latent common factor, and u is a vector of idiosyncratic shocks. The conditional density of the common factor f is assumed to be Gaussian, with first-order autoregressive dynamics and a switching unconditional mean,

$$P(f_t|h_t;\theta) \propto \exp\left[\frac{-1}{2\sigma^2}(f_t - \text{const}_{s_t} - \rho f_{t-1})^2\right], \tag{2.2}$$

where h contains past s, x, and f. Finally, the latent state takes the value $s = 0$ in expansions and $s = 1$ in recessions; its dynamics are first-order Markov, with transition dynamics

$$M = \begin{bmatrix} p_{00} & 1 - p_{00} \\ 1 - p_{11} & p_{11} \end{bmatrix}. \tag{2.3}$$

Generalizations could be entertained, such as incorporating time-varying transition probabilities as in Diebold, Lee, and Weinbach (1994), but the simple model (2.1)–(2.3) is well suited to our present needs.

3. THE EVIDENCE

The "small-data" variant of the regime-switching dynamic factor model has met with empirical success. Diebold and Rudebusch (1996) show that, although evidence of regime switching is hard to uncover in the four individual indicators that comprise the coincident index (i.e., common factor) extracted by Stock and Watson (1989), there is strong evidence of regime switching in the index itself. This is precisely what one expects in a regime-switching dynamic factor world, because the various individual indicators are contaminated by idiosyncratic noise, whereas the common factor is not.

The Diebold–Rudebusch (1996) approach is based on a two-step analysis in which one first extracts a small-data common factor by using the Kalman filter and then fits a Markov-switching model to the extracted factor. Simultaneous one-step estimation is difficult, because of the two levels of latency in the model: The common factor is latent, and the conditional dynamics of the factor are themselves dependent on a latent state. In important recent work, however, Kim and Nelson (1998, 1999) used Markov chain Monte Carlo methods to

Table 1. *Parameter estimates: Markov-switching model,* $\Delta \ln X\text{CI}$

$\hat{\mu}_0$	$\hat{\mu}_1$	$\hat{\rho}$	$\hat{\sigma}$	$\hat{p}_{00}$	$\hat{p}_{11}$
0.18	−1.24	0.22	0.78	0.97	0.80
(3.9)	(−7.2)	(4.5)	(26.6)	(11.8)	(3.0)

Note: The table contains maximum likelihood estimates of the parameters of the Markov-switching model given by (2.1)–(2.3) in the text, with t statistics in parentheses. The variable modeled is $\Delta \ln X\text{CI}$ (standardized to have zero mean and unit variance), where XCI is the Stock–Watson (1989) coincident index. The sample period is 1959:4–1998:12. See text for details.

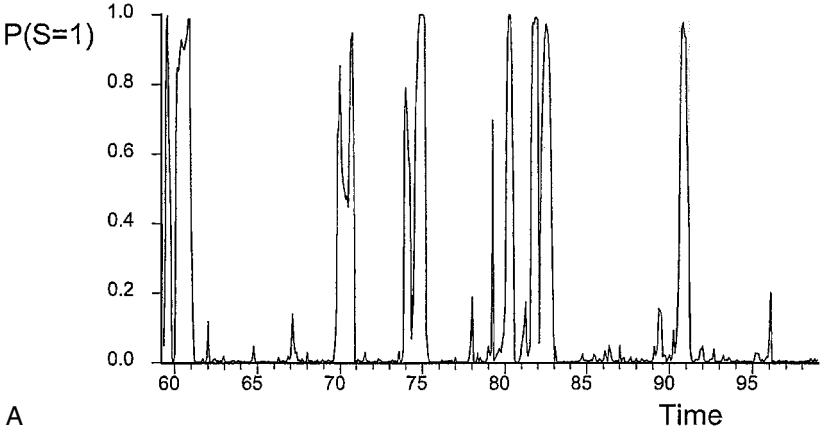
perform a one-step maximum likelihood (as well as Bayesian) estimation of the model, confirming and extending the earlier results.

To illustrate and confirm the evidence for regime switching in small-data dynamic factor models for macroeconomic analysis, I fit the Markov-switching model (2.1)–(2.3) to a common factor extracted from a very small number of indicators using the Stock–Watson (1989) methodology (i.e., their coincident index, XCI, obtained from the National Bureau of Economic Research (NBER) web page). The sample period is 1959:4–1998:12, and the variable modeled is $\Delta \ln X\text{CI}$, standardized to have zero mean and unit variance. The results appear in Table 1 and seem to clearly indicate switching across positive and negative growth regimes. The corresponding extracted recession probabilities, shown in Figure 1(a), show remarkable conformity to the shaded NBER recession chronology. The likelihood ratio test statistic for the null hypothesis of one state against the alternative of two is a large 29.5; despite the fact that it does not have a chi-square distribution (because of the nonidentification of nuisance parameters under the null, among other things), the value of 29.5 would likely remain highly significant even if Hansen's (1996) methods were used to generate corrected critical values, as indicated by the tabulations in Garcia (1998).

Now let us get back to Big Data. Call the first principal component extracted from Watson's Big Data PC1. It turns out that movements in PC1 and $\Delta \ln X\text{CI}$ (both standardized to have zero mean and unit variance) cohere very closely, as evidenced in the scatterplot in Figure 2 and the time-series plot in Figure 3. However, if PC1 is very close to $\Delta \ln X\text{CI}$, and $\Delta \ln X\text{CI}$ is well approximated by a Markov-switching process, then should not PC1 be as well? In Table 2, I show the results of fitting a Markov-switching model to PC1, and in Figure 1(b), I show the corresponding extracted recession probabilities. The evidence seems strongly in favor of switching; the likelihood ratio test statistic is 33.7.

4. CONCLUSIONS AND DIRECTIONS FOR FUTURE RESEARCH

The Reichlin and Watson papers, and the emerging Big Data dynamic factor modeling literature of which they are an important part, are major contributions



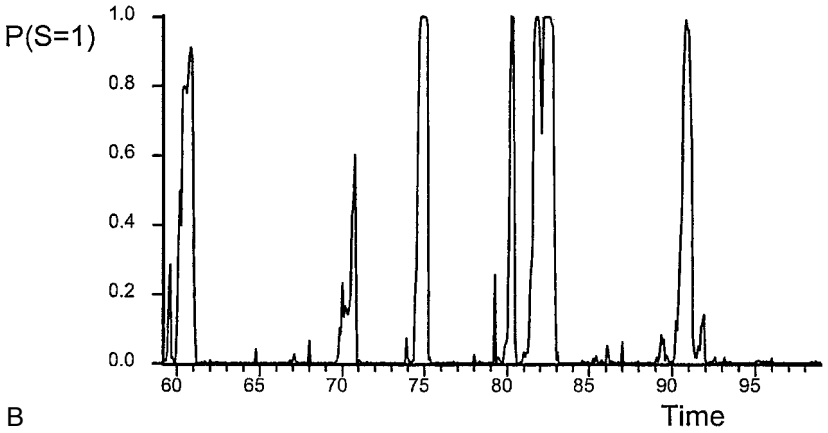


Figure 1. Smoothed recession probabilities: (a) $\Delta \ln XCI$ and (b) PC1.

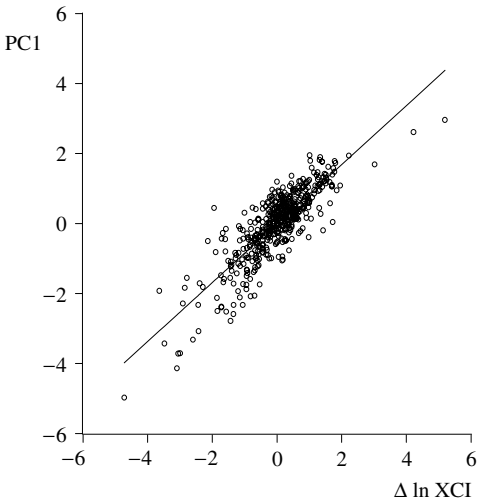


Figure 2. Scatter plot with fitted regression line (PC1 vs. $\Delta \ln XCI$).

Table 2. *Parameter estimates: Markov-switching model,* PC1

$\hat{\mu}_0$	$\hat{\mu}_1$	$\hat{\rho}$	$\hat{\sigma}$	$\hat{p}_{00}$	$\hat{p}_{11}$
0.12	−1.11	0.48	0.66	0.98	0.83
(3.2)	(−5.6)	(11.4)	(26.9)	(10.5)	(2.9)

Note: The table contains maximum likelihood estimates of the parameters of the Markov-switching model given by (2.1)–(2.3) in the text, with t statistics in parentheses. The variable modeled is Watson's first principal component, PC1 (standardized to have zero mean and unit variance). The sample period is 1959:4–1998:12. See text for details.

to empirical macroeconomics. They are, however, based on linear models and perspectives, whereas even the quick analyses performed here reveal a potentially important nonlinearity – regime switching – lurking just below the surface. Additional research in that direction will likely be fruitful.

Note well that regime switching does not necessarily invalidate the Big Data factor extractions of Reichlin and of Watson. Under conditions, their extractions are consistent for the factor even with regime switching (a fact that, by the way, could presumably be used to develop formal justification for the Diebold–Rudebusch two-step procedure). Hence regime switching need not be of central importance if, like Watson, one wants to use f to forecast y, because the factor is still extracted consistently and $E(y|f)$ may be linear even if the dynamics of f are nonlinear. But regime switching, if present, seems more unavoidably central in analyses such as Reichlin's, in which interest centers on the monitoring, interpreting, and forecasting of f.

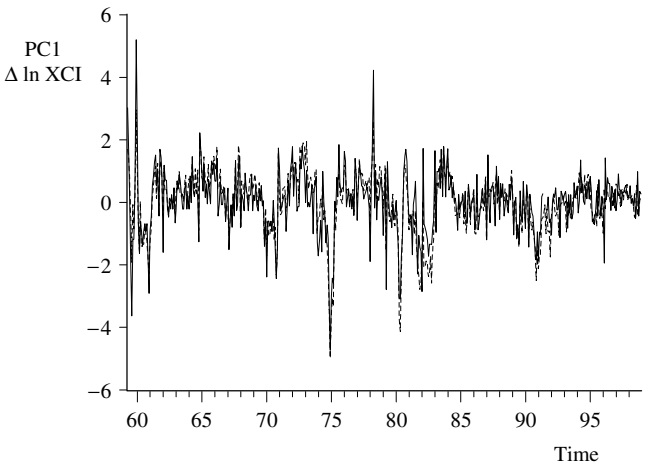


Figure 3. Time series plots (PC1 and $\Delta \ln X$CI).

The upshot is simply that, one way or another, the macroeconometric Big Data dynamic factor modeling literature should think harder about nonlinearity in general, and regime switching in particular. Allowing for regime switching in a dynamic factor model is one way of combining nonlinear with linear dynamics, and recent small-data work by Stock and Watson (1999) indicates that there may be forecasting gains from doing so, despite their finding that linear models dominate nonlinear models when one or the other is used alone. The regime-switching dynamic factor model is not the only way to combine nonlinear and linear dynamics, and it may not be the best way. But that is really the point: potentially important stones have been left unturned, and much is yet to be learned.

It seems appropriate to close with a forecast. Where are macroeconomic measurement and forecasting based on Big Data dynamic factor models headed? It is headed, and should be headed, toward calculation of recession probabilities conditional on the huge amount of data actually available (many thousands of series), in real time as the data are released, some measured quarterly, some monthly, some weekly, some such as asset prices in nearly continuous time, and some irregularly. And I conjecture – as is no surprise by now – that the recession probability calculations and forecasts will fruitfully be based on a model with a regime-switching factor. Finally, most of the work in the Big Data factor model literature has been likelihood based, as has all of this discussion, but I look forward to the formal blending of prior and likelihood information via Bayesian methods, and to assessing the robustness of posterior recession probabilities to alternative prior views. Both Reichlin and Watson are taking important and laudable steps in that direction.

ACKNOWLEDGMENTS

This note is a discussion of papers prepared for the World Congress of the Econometric Society, August, 2000. I am grateful to the National Science Foundation for support; to Rob Engle, Jim Stock, and Mark Watson for helpful comments; and to Sean Campbell for outstanding research assistance. Address correspondence to: F.X. Diebold, Department of Economics, University of Pennsylvania, 3718 Locust Walk, Philadelphia, PA 19104-6297. E-mail: fdiebold@mail.sas.upenn.edu.

References

Burns, A. F. and W. C. Mitchell (1946), *Measuring Business Cycles*. New York: National Bureau of Economic Research.

Diebold, F. X., J.-H. Lee, and G. C. Weinbach (1994), "Regime Switching with Time-Varying Transition Probabilities," in *Nonstationary Time-Series Analysis and Cointegration*, (ed. by C. Hargreaves), Oxford: Oxford University Press, 283–302.

(Reprinted in Diebold, F. X. and G. D. Rudebusch, 1999, *Business Cycles: Durations, Dynamics, and Forecasting*, Princeton, NJ: Princeton University Press.)

Diebold, F. X., and G. D. Rudebusch (1996), "Measuring Business Cycles: A Modern Perspective," *Review of Economics and Statistics*, 78, 67–77.

Garcia, R. (1998), "Asymptotic Null Distribution of the Likelihood Ratio Test in Markov Switching Models," *International Economic Review*, 39, 763–788.

Geweke, J. (1977), "The Dynamic Factor Analysis of Economic Time-Series Models," in *Latent Variables in Socioeconomic Models*, (ed. by D. J. Aigner and A. S. Goldberger), Amsterdam: North-Holland, 365–383.

Hamilton, J. D. (1989), "A New Approach to the Economic Analysis of Nonstationary Time Series and the Business Cycle," *Econometrica*, 57, 357–384.

Hansen, B. E. (1996), "Inference when a Nuisance Parameter Is Not Identified under the Null Hypothesis," *Econometrica*, 64, 416–430.

Kim, C.-J. and C. R. Nelson (1998), "Business Cycle Turning Points, A New Coincident Index, and Tests of Duration Dependence Based on a Dynamic Factor Model with Regime-Switching," *Review of Economics and Statistics*, 80, 188–201.

Kim, C.-J. and C. R. Nelson (1999), *State Space Models with Regime Switching*. Cambridge, MA: MIT Press.

Sargent, T. J. and C. Sims (1977), "Business Cycle Modeling Without Pretending to Have Too Much a priori Theory," in *New Methods of Business Cycle Research*, (ed. by C. Sims), Minneapolis: Federal Reserve Bank of Minneapolis.

Stock, J. H., and M. W. Watson (1989), "New Indexes of Coincident and Leading Economic Indicators," in *NBER Macroeconomics Annual*, (ed. by O. Blanchard and S. Fischer), Cambridge, MA: MIT Press, 351–394.

Stock, J. H., and M. W. Watson (1999), "A Comparison of Linear and Nonlinear Univariate Models for Forecasting Macroeconomic Time Series," in *Cointegration, Causality, and Forecasting: A Festschrift in Honor of Clive W. J. Granger*, (ed. by R. Engle and H. White), Oxford: Oxford University Press, 1–44.

CHAPTER 4

How Severe Is the Time-Inconsistency Problem in Monetary Policy?

Stefania Albanesi, V. V. Chari, and Lawrence J. Christiano

1. INTRODUCTION

The history of inflation in the United States and other countries has occasionally been quite bad. Are the bad experiences the consequence of policy errors? Or does the problem lie with the nature of monetary institutions? The second possibility has been explored in a long literature, which starts at least with Kydland and Prescott (1977) and Barro and Gordon (1983). This paper seeks to make a contribution to that literature.

The Kydland–Prescott and Barro–Gordon literature focuses on the extent to which monetary institutions allow policymakers to commit to future policies. A key result is that if policymakers cannot commit to future policies, inflation rates are higher than if they can commit. That is, there is a time-inconsistency problem that introduces a systematic inflation bias. This paper investigates the magnitude of the inflation bias in two standard general equilibrium models. One is the cash–credit good model of Lucas and Stokey (1983). The other is the limited-participation model of money described in Christiano, Eichenbaum, and Evans (1997). We find that, for a large range of parameter values, there is no inflation bias.

In the Kydland–Prescott and Barro–Gordon literature, equilibrium inflation in the absence of commitment is the outcome of an interplay between the benefits and costs of inflation. For the most part, this literature consists of reduced-form models. Our general equilibrium models[1] incorporate the kinds of benefits and costs that seem to motivate the reduced-form specifications. To understand these benefits and costs, we must first explain why money is not neutral in our models. In each case, at the time the monetary authority sets its money growth rate, some nominal variable in the economy has already been set. In the cash–credit good model, this variable is the price of a subset of intermediate goods. As in Blanchard and Kiyotaki (1987), some firms must post prices in advance and are required to meet all demand at their posted price. In

[1] See Chari, Christiano, and Eichenbaum (1998), Ireland (1997), and Neiss (1999) for related general equilibrium models.

the limited-participation model, a portfolio choice variable is set in advance. In each case, higher than expected money growth tends – other things the same – to raise output. The rise in output raises welfare because the presence of monopoly power in our model economies implies that output and employment are below their efficient levels. These features give incentives to the monetary authority to make money growth rates higher than expected.

Turning to the costs of inflation, we first discuss the cash–credit good model. We assume that cash good consumption must be financed by using money carried over from the previous period.[2] If the money growth rate is high, the price of the cash good is high and the quantity of cash goods consumed is low. This mechanism tends to reduce welfare as the money growth rate rises. The monetary authority balances the output-increasing benefits of high money growth against the costs of the resulting fall in cash good consumption. Somewhat surprisingly, we find that there is a large subset of parameter values in which the costs of inflation dominate the benefits at all levels of inflation and money growth above the ex ante optimal rate. As a result, for these parameter values, the unique equilibrium yields the same outcome as under commitment.

In our limited-participation model, at all interest rates higher than zero, increases in money growth tend to stimulate employment by reducing the interest rate. As a result, there is no equilibrium with a positive interest rate. When the interest rate is already zero, further reductions are not possible. In this case, additional money generated by the monetary authority simply accumulates as idle balances at the financial intermediary. The unique Markov equilibrium in this model has a zero interest rate. Again, there is no time-inconsistency problem and no inflation bias.

Should we conclude from our examples that lack of commitment in monetary policy cannot account for the bad inflation outcomes that have occurred? We think such a conclusion is premature. Research on the consequences of lack of commitment in dynamic general equilibrium models is still in its infancy. Elsewhere, in Albanesi, Chari, and Christiano (2002), we have displayed a class of empirically plausible models in which lack of commitment may in fact lead to high and volatile inflation. The key difference between the model in that paper and the models studied here lies in the modeling of money demand. Taken together, these findings suggest that a resolution of the importance of time inconsistency in monetary policy depends on the details of money demand. As our understanding about the implications for time inconsistency in dynamic models grows, we may discover other features of the economic environment that are crucial for determining the severity of the time-inconsistency problem. It is too soon to tell whether the ultimate conclusion will be consistent with the implications of the models studied in this paper.

The paper is organized as follows. In Section 2, we analyze a cash–credit goods model with arbitrary monetary policies. This section sets up the basic

[2] We assume a timing structure as in Svensson (1985) rather than as in Lucas and Stokey (1983). See also Nicolini (1998).

framework for analyzing purposeful monetary policy. Interestingly, we also obtain some new results on multiplicity of equilibria under mild deflations. In Section 3 we analyze the model in Section 2 when monetary policy is chosen by a benevolent policy maker without commitment. In Section 4, we analyze a limited-participation model. In Section 5 we give our conclusions.

2. EQUILIBRIUM IN A CASH–CREDIT GOOD MODEL WITH ARBITRARY MONETARY POLICY

Here we develop a version of the Lucas–Stokey cash–credit good model. There are three key modifications: we introduce monopolistic competition, as in Blanchard and Kiyotaki (1987); we modify the timing in the cash-in-advance constraint as indicated in the Introduction; and we consider nonstationary equilibria. The agents in the model are a representative household and representative intermediate and final good producers. A policy for the monetary authority is a sequence of growth rates for the money supply. We consider arbitrary monetary policies and define and characterize the equilibrium. We show that in the best equilibrium with commitment, monetary policy follows the Friedman rule in the sense that the nominal interest rate is zero. Following Cole and Kocherlakota (1998), we show that there is a nontrivial class of monetary policies that support the best equilibrium. Interestingly, we show that only one of the policies in this class is robust, whereas the others are fragile. Specifically, we show that only the policy in which money growth deflates at the pure rate of time preference supports the best equilibrium as the unique outcome. We show that the other policies are fragile in the sense that there are many equilibria associated with them.

2.1. Households

The household's utility function is

$$\sum_{t=0}^{\infty} \beta^t u(c_{1t}, c_{2t}, n_t), \quad u(c_1, c_2, n) = \log c_{1t} + \log c_{2t} + \log(1 - n), \tag{2.1}$$

where c_{1t}, c_{2t}, and n_t denote consumption of cash goods, consumption of credit goods, and employment, respectively.

The sequence of events in the period is as follows. At the beginning of the period, the household trades in a securities market in which it allocates nominal assets between money and bonds. After trading in the securities market, the household supplies labor and consumes cash and credit goods.

For securities market trading, the constraint is

$$A_t \geq M_t + B_t, \tag{2.2}$$

where A_t denotes beginning-of-period t nominal assets, M_t denotes the household's holdings of cash, B_t denotes the household's holdings of interest-bearing

bonds, and A_0 is given. Cash goods must be paid for with currency from securities market trading. The cash-in-advance constraint is given by

$$P_{1t}c_{1t} \leq M_t. \tag{2.3}$$

The household's sources of cash during securities market trading are cash left over from the previous period's goods market, $M_{t-1} - P_{1t-1}c_{1t-1}$, earnings on bonds accumulated in the previous period, $R_{t-1}B_{t-1}$, transfers received from the monetary authority, T_{t-1}, labor income in the previous period, $W_{t-1}n_{t-1}$, and profits in the previous period, D_{t-1}. Let P_{1t-1}, P_{2t-1}, and R_{t-1} denote the period $t-1$ prices of cash and credit goods, and the gross interest rate, respectively. Finally, the household pays debts, $P_{2t-1}c_{2t-1}$, owed from its period $t-1$ purchases of credit goods during securities market trading. These considerations are summarized in the following securities market constraint:

$$\begin{aligned} A_t = {} & W_{t-1}n_{t-1} - P_{2t-1}c_{2t-1} + (M_{t-1} - P_{1t-1}c_{1t-1}) \\ & + R_{t-1}B_{t-1} + T_{t-1} + D_{t-1}. \end{aligned} \tag{2.4}$$

We place the following restriction on the household's ability to borrow:

$$A_{t+1} \geq -\frac{1}{q_{t+1}} \sum_{j=1}^{\infty} q_{t+j+1}\,[W_{t+j} + T_{t+j} + D_{t+j}], \quad \text{for } t = 0, 1, 2, \ldots, \tag{2.5}$$

where $q_t = \prod_{j=0}^{t-1} 1/R_j$, $q_0 \equiv 1$. Condition (2.5) says that the household can never borrow more than the maximum present value future income.

The household's problem is to maximize (2.1) subject to (2.5)–(2.2) and the nonnegativity constraints, n_t, c_{1t}, c_{2t}, and $1 - n_t \geq 0$. If $R_t < 1$ for any t, this problem does not have a solution. We assume throughout that $R_t \geq 1$.

2.2. Firms

We adopt a variant of the production framework in Blanchard and Kiyotaki (1987). In developing firm problems, we delete the time subscript. In each period, there are two types of perfectly competitive, final goods firms: those that produce cash goods and those that produce credit goods. Their production functions are

$$y_1 = \left[\int_0^1 y_1(\omega)^{\lambda} d\omega\right]^{1/\lambda}, \quad y_2 = \left[\int_0^1 y_2(\omega)^{\lambda} d\omega\right]^{1/\lambda}, \tag{2.6}$$

where y_1 denotes output of the cash good, y_2 denotes output of the credit good, and $y_i(\omega)$ is the quantity of intermediate good of type ω used to produce good i, and $0 < \lambda < 1$. These firms solve

$$\max_{y_i, \{y_i(\omega)\}} P_i y_i - \int_0^1 P_i(\omega) y_i(\omega) d\omega, \quad i = 1, 2.$$

Solving this problem leads to the following demand curves for each intermediate

good:

$$y_i(\omega) = y_i \left[\frac{P_i}{P_i(\omega)} \right]^{1/(1-\lambda)}, \quad i = 1, 2. \tag{2.7}$$

Intermediate good firms are monopolists in the product market and competitors in the market for labor. They set prices for their goods and are then required to supply whatever final good producers demand at those prices. The intermediate good firms solve

$$\max_{y_i(\omega)} P_i(\omega) y_i(\omega) - W n_i(\omega), \quad i = 1, 2,$$

where W is the wage rate, subject to a production technology, $y_i(\omega) = n_i(\omega)$, and the demand curve in (2.7). Profit maximization leads the intermediate good firms to set prices according to a markup over marginal costs:

$$P_1(\omega) = \frac{W}{\lambda}, \quad P_2(\omega) = \frac{W}{\lambda}. \tag{2.8}$$

2.3. Monetary Authority

At date t, the monetary authority transfers T_t units of cash to the representative household. It finances the transfers by printing money. Let g_t denote the growth rate of the money supply. Then, $T_t = (g_t - 1)M_t$, where M_0 is given and $M_{t+1} = g_t M_t$. A monetary policy is an infinite sequence, g_t, where $t = 0, 1, 2, \ldots$.

2.4. Equilibrium

We begin by defining an equilibrium, given an arbitrary specification of monetary policy. We then discuss the best equilibrium achievable by some monetary policy. This equilibrium is one in which the nominal interest rate is zero. Thus, the Friedman rule is optimal in this model. We go on to discuss the set of policies that support the best equilibrium.

Definition 2.1. *A private sector equilibrium is a set of sequences,* $\{P_{1t}, P_{2t}, W_t, R_t, c_{1t}, c_{2t}, n_t, B_t, M_t, g_t\}$, *with the following properties:*

1. *Given the prices and the government policies, the quantities solve the household problem.*
2. *The firm optimality conditions in (2.8) hold.*
3. *The various market-clearing conditions hold:*

$$c_{1t} + c_{2t} = n_t, \quad B_t = 0, \quad M_{t+1} = M_t g_t. \tag{2.9}$$

Next, we define the best equilibrium.

Definition 2.2. *A Ramsey equilibrium is a private sector equilibrium with the highest level of utility.*

We now develop a set of equations that, together with (2.7)–(2.9), allow us to characterize a private sector equilibrium. From (2.8) it follows that $P_{1t} = P_{2t}$. Let $P_t = P_{1t} = P_{2t}$. Combining the household and the firm first-order conditions, we get

$$\frac{c_{2t}}{1 - c_{1t} - c_{2t}} = \lambda, \quad \text{all } t. \tag{2.10}$$

$$P_{t+1} c_{1t+1} = \beta R_t P_t c_{1t}, \quad \text{all } t. \tag{2.11}$$

$$R_t = \frac{c_{2t}}{c_{1t}} \geq 1, \quad \text{all } t. \tag{2.12}$$

$$P_t c_{1t} - M_t \leq 0, \quad (R_t - 1)(P_t c_{1t} - M_t) = 0. \tag{2.13}$$

In equilibrium, with $B_t = 0$, the household's transversality condition is

$$\lim_{t \to \infty} \beta^t \frac{M_t}{P_t c_{1t}} = 0. \tag{2.14}$$

The nonnegativity constraint on leisure implies

$$c_{1t} + c_{2t} \leq 1. \tag{2.15}$$

We summarize these results in the form of a proposition.

Proposition 2.3 (characterization result). *A sequence,* $\{P_{1t}, P_{2t}, Wt, R_t, c_{1t}, c_{2t}, n_t, B_t, M_t, g_t\}$, *is an equilibrium if and only if (2.8)–(2.15) and* $P_{1t} = P_{2t} = P_t$ *are satisfied. Furthermore, for any* $R_t \geq 1$, *there exists a private sector equilibrium with employment and consumption allocations uniquely determined by*

$$n_{1t} = c_{1t} = \frac{\lambda}{\lambda + (1 + \lambda) R_t}, \quad n_{2t} = c_{2t} = R_t c_{1t}, \quad \text{all } t. \tag{2.16}$$

Proof. Equations (2.10)–(2.15) are the resource constraints and the necessary and sufficient conditions for household and firm optimization. Necessity and sufficiency in the case of the firms are obvious, and in the case of the households, the results are derived formally in Appendix A.

We now turn to the second part of the proposition. We need to verify that prices and a monetary policy can be found such that, together with the given sequence of interest rates and (2.16), they constitute a private sector equilibrium. First, by construction of (2.16), it can be verified that (2.9), (2.10), and (2.12) are satisfied. It can also be verified that (2.15) is satisfied. Second, let $P_0 = M_0 / c_{1,0}$, and use this and (2.11) to compute P_t, for $t = 1, 2, 3, \ldots$. This construction ensures that (2.11) for all t and (2.13) for $t = 0$ are satisfied. Next, we compute $M_t = P_t c_{1t}$ for $t = 1, 2, \ldots$, so that (2.13) is satisfied for all t. Finally, (2.14) is satisfied because $0 < \beta < 1$ and $M_t / (P_t c_{1t}) = 1$. ■

We use this proposition to characterize the Ramsey equilibrium:

Proposition 2.4 (Ramsey equilibrium yields Friedman rule). *Any Ramsey equilibrium has the property $R_t = 1$ for all t and employment and consumption allocations given in (2.16).*

Proof. The Ramsey equilibrium solves

$$\max_{\{R_t \geq 1\}_{t=0}^{\infty}} \sum_{t=0}^{\infty} \beta^t \{2 \log(c_{1t}) + \log R_t + \log[1 - (1 + R_t)c_{1t}]\},$$

where c_{1t} is given by (2.16). This problem is equivalent to the static problem $\max_{R \geq 1} f(R)$, where $f(R) = 2 \log(c_1) + \log R + \log[1 - (1 + R)c_1]$, $c_1 = \lambda/[\lambda + (1 + \lambda)R]$. This function is concave in R and is maximized at the corner solution $R = 1$. ■

We now turn to the set of policies that are associated with a Ramsey equilibrium. The next proposition shows that there is a continuum of such policies. It is the analog of Proposition 2 in Cole and Kocherlakota (1998).

Proposition 2.5 (policies associated with Ramsey equilibrium). *There exists a private sector equilibrium with $R_t = 1$ for all t if and only if*

$$\frac{M_t}{\beta^t} \geq \kappa, \quad \kappa > 0 \quad \text{for all } t \tag{2.17}$$

$$\lim_{T \to \infty} M_T \to 0. \tag{2.18}$$

Proof. Consider the necessity of (2.17) and (2.18). Suppose we have an equilibrium satisfying $R_t = 1$ and (2.8)–(2.15). From (2.12) and (2.10), letting $c_t \equiv c_{1t} = c_{2t}$, we obtain

$$c_t = c = \frac{\lambda}{1 + 2\lambda} \quad \text{for all } t. \tag{2.19}$$

From (2.11) we obtain

$$P_t c_t = \beta^t P_0 c > 0. \tag{2.20}$$

Substituting (2.20) into (2.14), we get

$$\lim_{t \to \infty} \beta^t \frac{M_t}{P_t c_t} = \lim_{t \to \infty} \beta^t \frac{M_t}{\beta^t P_0 c} = 0,$$

so that (2.18) is satisfied. From the cash-in-advance constraint in (2.13),

$$\beta^t P_0 c \leq M_t \quad \text{for each } t,$$

which implies (2.17).

Consider sufficiency. Suppose (2.17) and (2.18) are satisfied, and that $R_t = 1$. We must verify that the other nonzero prices and quantities can be found that satisfy (2.8)–(2.15). Let $c_{1t} = c_{2t} = c$ in (2.19) for all t. Let $P_t = \beta^t P_0$, where $P_0 > 0$ will be specified in the paragraphs that follow. These two specifications

guarantee (2.10), (2.12), and (2.11). Condition (2.18), together with the given specification of prices and consumption, guarantees (2.14). Finally, it is easily verified that when $0 < P_0 \leq \kappa/c$ is set, the cash-in-advance constraint in (2.13) holds for each t. ■

The previous proposition shows that there are many policies that implement the Ramsey outcome. However, many of these policies are fragile in the sense that they can yield worse outcomes than the Ramsey outcome. The next proposition characterizes the set of equilibria associated with mild monetary deflations in which the (stationary) growth rate of the money supply satisfies $\beta < g < 1$.

Proposition 2.6 (fragility of mild monetary deflations). *If $\beta < g < 1$, the following are equilibrium outcomes:*

(i) $R_t = 1, c_{1t} = c_{2t} = \lambda/[1 + 2\lambda]$ *for all* $t, P_{t+1}/P_t = \beta, M_t/P_t \to \infty$.

(ii) $R_t = g/\beta, c_{1t} = \lambda/[\lambda + (1 + \lambda)g/\beta], c_{2t} = (g/\beta)c_{1t}, P_{t+1}/P_t = g$ *for all* $t, M_t/P_t$ *independent of* t.

(iii) $R_t = g/\beta$ *for* $t \leq t^*$, $R_t = 1$ *for* $t > t^*$ *for* $t^* = 0, 1, 2, \ldots, c_{1t} = \lambda/[\lambda + (1 + \lambda)R_t], c_{2t} = R_t c_{1t}$,

$$\frac{P_{t+1}}{P_t} = \begin{cases} g, t = 0, 1, \ldots, t^* - 1 (\text{for } t^* > 0) \\ \dfrac{(1 + 2\lambda)g}{\lambda + (g/\beta)(1 + \lambda)}, t = t^*. \\ \beta, t = t^* + 1, t^* + 1, \ldots \end{cases}$$

Proof. That these are all equilibria may be confirmed by verifying that (2.8)–(2.15) are satisfied. ■

This proposition does not characterize the entire set of equilibria that can occur with $\beta < g < 1$. It gives a flavor of the possibilities, however. For example, (iii) indicates that there is a countable set of equilibria (one for each possible t^*) in which the consumption and employment allocations are not constant and the interest rate switches down to unity after some date. Although there do exist equilibria in which consumption and employment are not constant, they appear to be limited. For example, it can be shown that there is no equilibrium in which the interest rate switches up from unity at some date; that is, there does not exist an equilibrium in which $R_{t^*} = 1$ and $R_{t^*+1} > 1$ for some t^*. To see this, suppose the contrary. Then, from (2.11), $\beta P_{t^*} c_{1t^*} = P_{t^*+1} c_{1t^*+1} = M_{t^*+1}$, because the cash-in-advance constraint must be binding in period $t^* + 1$. However, $M_{t^*} \geq P_{t^*} c_{1t^*}$ implies $\beta \geq g$, a contradiction. Furthermore, we can also show that there do not exist equilibria in which the interest rate changes and it is always greater than unity, that is, in which $R_t \neq R_{t+1} \neq 1$. So, although the set of equilibria with nonconstant interest rates (and, hence, nonconstant consumption) is limited, Proposition 2.6 indicates that it does exist.

The preceding proposition indicates that mild monetary deflations are fragile. It turns out, however, that a deflationary policy of the kind advocated by Milton Friedman is robust in the sense that it always yields the Ramsey outcome.

Proposition 2.7 (robustness of Friedman deflation). *Suppose $g_t = \beta$. Then, all equilibria are Ramsey equilibria.*

Proof. To show that if $g_t = \beta$, $R_t = 1$, suppose the contrary. That is, $R_t > 1$ for some t. Therefore, $P_t c_{1t} = M_t$. Also, $P_{t+1} c_{1t+1} \leq M_{t+1}$. By (2.11) we find $1/(P_t c_{1t}) = \beta R_t/(P_{t+1} c_{1t+1})$, so that $(1/M_t) \geq \beta R_t (1/M_{t+1})$, or $g_t \geq \beta R_t$, which is a contradiction. ■

It is worth pointing out that because the interest rate is constant, so are real allocations. There are, however, a continuum of equilibria in which the price level is different. In all of these equilibria, $P_{t+1}/P_t = \beta$. These equilibria are indexed by the initial price level, P_0, which satisfies $P_0 \leq M_0[1 + 2\lambda]/\lambda$ and $P_{t+1}/P_t = \beta$.

3. MARKOV EQUILIBRIUM IN A CASH–CREDIT GOOD MODEL

In this section we analyze a version of the model presented herein in which a benevolent government chooses monetary policy optimally. We consider a more general utility function of the constant elasticity of substitution (CES) form:

$$u(c_1, c_2, n) = \frac{1}{1-\sigma}\left[\left(\alpha c_1^{\rho} + (1-\alpha)c_2^{\rho}\right)^{1/\rho}(1-n)^{\gamma}\right]^{1-\sigma}.$$

Note that this utility function is a generalization of the one used in the previous section. Here, we focus on the Markov equilibrium of this model. The timing is as follows. A fraction, μ_1, of intermediate good producers in the cash good sector and a fraction, μ_2, of intermediate good producers in the credit good sector set prices at the beginning of the period. These firms are referred to as sticky price firms. We show in what follows that all sticky price firms set the same price. Denote this price by P^e. This price, all other prices, and all nominal assets in this section are scaled by the aggregate, beginning-of-period money stock. Then, the monetary authority chooses the growth rate of the money supply. Finally, all other decisions are made.

The state of the economy at the time the monetary authority makes its decision is P^e.[3] The monetary authority makes its money growth decision conditional on P^e. We denote the gross money growth rate by G and the policy rule by $X(P^e)$. The state of the economy after the monetary authority makes

[3] Notice that we do not include the aggregate stock of money in the state. In our economy, all equilibria are neutral in the usual sense that if the initial money stock is doubled, there is an equilibrium in which real allocations and the interest rate are unaffected and all nominal variables are doubled. This consideration leads us to focus on equilibria that are invariant with respect to the initial money stock. We are certainly mindful of the possibility that there can be equilibria that depend on the money stock. For example, if there are multiple equilibria in our sense, it is possible to construct "trigger strategy-type" equilibria that are functions of the initial money stock. In our analysis, we exclude such equilibria, and we normalize the aggregate stock of money at the beginning of each period to unity.

its decision is $S = (P^e, G)$. With these definitions of the economy's state variables, we proceed now to discuss the decisions of firms, households, and the monetary authority.

Recall that profit maximization leads intermediate good firms to set prices as a markup over the wage rate; see Equation (2.8). The price set by the $1 - \mu_1$ intermediate good producers in the cash good sector and $1 - \mu_2$ intermediate good firms in the credit good sector that set their prices after the monetary authority makes its decision ("flexible price firms") is denoted by $\hat{P}(S)$. For the μ_1 and μ_2 sticky price cash and credit good firms, respectively, and the $1 - \mu_1$ and $1 - \mu_2$ flexible price cash and credit good firms, respectively, the markup rule implies

$$\begin{aligned} P^e &= \frac{W[P^e, X(P^e)]}{\lambda}, \\ \hat{P}(S) &= \frac{W(S)}{\lambda}, \quad 0 < \lambda < 1, \end{aligned} \tag{3.1}$$

where $W(S)$ denotes the nominal wage rate. In this model of monopolistic competition, output and employment are demand determined. That is, output and employment are given by (2.8). Let $P_i(S)$ denote the price of the cash and credit good for $i = 1, 2$, respectively. Let $y_{ij}(S)$, where $i, j = 1, 2$, denote the output of the intermediate goods firms, where the first subscript denotes whether the good is a cash good ($i = 1$) or a credit good ($i = 2$), and the second subscript indicates whether it is produced by a sticky price ($j = 1$) or a flexible price ($j = 2$) producer.

In terms of the household's problem, it is convenient to write the constraints in recursive form. The analog of (2.2) is

$$M + B \leq A, \tag{3.2}$$

where, recall, A denotes beginning-of-period nominal assets, M denotes the household's holdings of cash, and B denotes the household's holdings of interest-bearing bonds. Here, nominal assets, money, and bonds are all scaled by the aggregate stock of money. We impose a no-Ponzi constraint of the form $B \leq \bar{B}$, where $\bar{B}$ is a large, finite, upper bound. The household's cash-in-advance constraint is

$$M - P_1(S)c_1 \geq 0, \tag{3.3}$$

where c_1 denotes the quantity of the cash good. Nominal assets evolve over time as follows:

$$\begin{aligned} 0 \leq\ & W(S)n + [1 - R(S)]M - P_1(S)c_1 - P_2(S)c_2 \\ & + R(S)A + (G - 1) + D(S) - GA', \end{aligned} \tag{3.4}$$

where c_2 denotes the quantities of credit goods purchased. In (3.4), $R(S)$ denotes

the gross nominal rate of return on bonds, and $D(S)$ denotes profits after lump sum taxes. Finally, B has been substituted out in the asset equation using (3.2). Notice that A' is multiplied by G. This modification is necessary because of the way we have scaled the stock of nominal assets.

Consider the household's asset, goods, and labor market decisions. Given that the household expects the monetary authority to choose policy according to X in the future, the household solves the following problem:

$$v(A, S) = \max_{n, M, A', c_i; i=1,2} u(c_1, c_2, n) + \beta v[A', P^e, X(P^e)], \tag{3.5}$$

subject to (3.2), (3.3), (3.4), and nonnegativity on allocations. In (3.5), v is the household's value function. The solution to (3.5) yields decision rules of the form $n(A, S)$, $M(A, S)$, $A'(A, S)$, and $c_i(A, S)$, where $i = 1, 2$. We refer to these decision rules, together with the production decisions of firms, $y_{ij}(S)$, $i, j = 1, 2$, as *private sector allocation rules*. We refer to the collection of prices, P^e, $\hat{P}(S)$, $W(S)$, $R(S)$, and $[P_i(S), i = 1, 2]$, as *pricing rules*.

3.1. Monetary Authority

The monetary authority chooses the current money growth rate, G, to solve the problem

$$\max_G v(1, S), \tag{3.6}$$

where, recall, $S = (P^e, G)$. Let $X(P^e)$ denote the solution to this problem. We refer to this solution as the *monetary policy rule*.

3.2. Markov Equilibrium

We now define a Markov equilibrium. This equilibrium requires that households and firms optimize and markets clear.

Definition 3.1. *A Markov equilibrium is a set of private sector allocation rules, pricing rules, a monetary policy rule, and a value function for households such that:*

(i) The value function, v, and the private sector rules solve (3.5).

(ii) Intermediate good firms optimize; that is, (3.1) is satisfied, final good prices satisfy

$$P_i(S) = \left[\mu_i (P^e)^{\lambda/(\lambda-1)} + (1 - \mu_i)\widehat{P}(S)^{\lambda/(\lambda-1)}\right]^{(\lambda-1)/\lambda}, \quad \text{for } i = 1, 2,$$

and the output of intermediate good firms, $y_{ij}(S)$, is given by the analog of (2.7).

(iii) Asset markets clear, that is, $A'(1, S) = 1$ *and* $M(1, S) = 1$.

(iv) The labor market clears, that is,

$$n(1, S) = \mu_1 y_{11}(S) + (1 - \mu_1) y_{12}(S) + \mu_2 y_{21}(S) + (1 - \mu_2) y_{22}(S).$$

(v) The monetary authority optimizes; that is, $X(P^e)$ *solves (3.6).*

Notice that our notion of Markov equilibrium has built into it the idea of sequential optimality captured in game-theoretic models by subgame perfection. In particular, we require that for any deviation by the monetary authority from $X(P^e)$, the resulting allocations be the ones that would actually occur, that is, the ones that would be in the best interests of households and firms and would clear markets.

We now define a Markov equilibrium outcome:

Definition 3.2. *A Markov equilibrium outcome is a set of numbers,* n, c_1, c_2, y_{ij} $(i, j = 1, 2)$, P^e, W, R, P_1, P_2, *and* g, *satisfying* $n = n(1, P^e, g), c_1 = c_1(1, P^e, g), \ldots,$ *and* $g = X(P^e)$.

3.3. Analysis of the Markov Equilibrium

Here we characterize the Markov equilibrium. In particular, we provide sufficient conditions for the Ramsey outcomes to be Markov equilibrium outcomes. We also provide sufficient conditions for the Markov equilibrium to be unique. Combining these conditions, we obtain sufficient conditions for the unique Markov equilibrium to yield the Ramsey outcomes.

In developing these results, we find it convenient to recast the monetary authority's problem as choosing $\hat{P}$ rather than G. First, we analyze the private sector allocation rules and pricing functions. Then, we analyze the monetary authority's problem.

We use the necessary and sufficient conditions of private sector maximization and market clearing to generate the private sector allocation rules and pricing functions. The conditions are given by the following:

$$-\frac{u_3}{u_2} = \lambda \frac{\hat{P}}{P_2}, \tag{3.7}$$

$$\left(\frac{1}{P_1} - c_1\right)(R - 1) = 0, \tag{3.8}$$

$$R = \frac{u_1}{u_2}\frac{P_2}{P_1}, \tag{3.9}$$

$$n_i = c_i \left[\mu_i \left(\frac{P_i}{P^e}\right)^{1/(1-\lambda)} + (1 - \mu_i)\left(\frac{P_i}{\hat{P}}\right)^{1/(1-\lambda)}\right], \quad \text{for } i = 1, 2, \tag{3.10}$$

$$n = n_1 + n_2, \tag{3.11}$$

$$P_i = \left[\mu_i (P^e)^{\lambda/(\lambda-1)} + (1-\mu_i)\hat{P}^{\lambda/(\lambda-1)}\right]^{(\lambda-1)/\lambda}, \quad \text{for } i = 1, 2, \tag{3.12}$$

$$\frac{Gu_1}{P_1} = \beta R v_1[1, P^e, X(P^e)]. \tag{3.13}$$

Notice that the growth rate of the money supply, G, appears only in (3.13). Equations (3.7)–(3.12) constitute eight equations in the eight unknowns, $c_1, c_2, n_1, n_2, n, P_1, P_2$, and R. Given values for P^e and $\hat{P}$, we see that these equations can be solved to yield functions of the following form:

$$c_1(P^e, \hat{P}), c_2(P^e, \hat{P}), \ldots, R(P^e, \hat{P}). \tag{3.14}$$

Replacing $\hat{P}$ in (3.14) by a pricing function, $\hat{P}(P^e, G)$, we obtain the allocation rules and pricing functions in a Markov equilibrium.

The pricing function, $\hat{P}(P^e, G)$, is obtained from Equation (3.13). This equation can be thought of as yielding a function, $G(P^e, \hat{P})$. The pricing function, $\hat{P}(P^e, G)$, is obtained by inverting $G(P^e, \hat{P})$. It is possible that the inverse of $G(P^e, \hat{P})$ is a correspondence. In this case, $\hat{P}(P^e, G)$ is a selection from the correspondence. Any such selection implies a range of equilibrium prices, $\hat{P}$. Denote this range by D.

Given the function, $\hat{P}(P^e, G)$, the monetary authority's problem can be thought of in either of two equivalent ways: either it chooses G or it chooses $\hat{P}$. The government's decision problem is simplified in our setting because its choice of $\hat{P}$ has no impact on future allocations. As a result, the government faces a static problem.

The allocation functions in (3.14) can be substituted into the utility function to obtain the following:

$$U(P^e, \hat{P}) = u[c_1(P^e, \hat{P}), c_2(P^e, \hat{P}), n(P^e, \hat{P})]. \tag{3.15}$$

Then, define

$$P(P^e) = \arg\max_{\hat{P} \in D} U(P^e, \hat{P}).$$

The function, $P(P^e)$, is the monetary authority's best response, given P^e. Equilibrium requires that $P(P^e) = P^e$. This procedure determines the expected price P^e, the actual price $\hat{P}$, and the eight allocations and other prices just described. Given these values, we can determine the equilibrium growth rate of the money supply by evaluating $G(P^e, P^e)$.

In what follows, we assume that the first-order conditions to the monetary authority's problem characterize a maximum. In quantitative exercises we have done using these models, we have found that the first-order conditions in the neighborhood of a Ramsey outcome do in fact characterize the global maximum of the monetary authority's problem.

Next, we show that for a class of economies the Ramsey outcomes are Markov equilibrium outcomes. Recall that a Ramsey equilibrium is a private sector equilibrium with $R = 1$. In Appendix A, we prove the following result:

Proposition 3.3 (Markov is Ramsey). *Suppose*

$$(1-\rho)(1-\mu_1) \geq \mu_2 \left(\frac{1-\alpha}{\alpha}\right)^{1/(1-\rho)}.$$

Then, there exists a Markov equilibrium with $R = 1$.

The intuition for this proposition is as follows. A benefit of expansionary monetary policy is that it leads to an increase in demand for goods whose prices are fixed. This increase in demand tends to raise employment. Other things being the same, welfare rises because employment is inefficiently low. A principal cost of expansionary monetary policy is that it tends to reduce employment in the cash good sector. The reason for this reduction in employment is that nominal consumption of the cash good is predetermined, while its price rises as a result of the increase in flexible intermediate good prices. It is possible that the reduction in employment in the cash good sector is so large that overall employment and welfare fall. Indeed, it can be shown that if the sufficient condition of the proposition is met, employment falls with an increase in the money growth rate in the neighborhood of the Ramsey equilibrium. The monetary authority has an incentive to contract the money supply. This incentive disappears only if the nominal interest rate is zero.

In what follows, we assume that $\hat{P}(P^e, G)$ is a continuous function of G. This restriction is not innocuous. We have constructed examples where, for a given value of G, there is more than one value of private sector allocations and prices that satisfy the conditions for private sector optimization and market clearing.[4] Thus, it is possible to construct private sector allocation rules and pricing functions that are discontinuous functions of G. The assumption of continuity plays an important role in the proof of uniqueness given here. In the next proposition we provide sufficient conditions for uniqueness of the Markov equilibrium:

Proposition 3.4 (uniqueness of Markov equilibrium). *Suppose:*

(i) $\rho = 0, \sigma = 1$.
(ii) $1 - \mu_1 > \mu_2[(1-\alpha)/\alpha]$.
(iii) $\{\lambda + [\gamma\alpha/(1-\alpha)]\}(1-\mu_1) \geq [(1-\lambda)\gamma\mu_2]/\alpha$.

Then, in the class of Markov equilibria in which $\hat{P}(P^e, G)$ *is a continuous function of G, (a) there exists an equilibrium with* $R = 1$ *and (b) there is no equilibrium with outcome* $R > 1$.

[4] Specifically, we found numerical examples in which the function, $G(P^e, \hat{P})$, displayed an inverted U shape when graphed for fixed P^e with G on the vertical axis and $\hat{P}$ on the horizontal. In these examples, each fixed $\hat{P}$ implied a unique G. However, there are intervals of values of G where a fixed G maps into two distinct $\hat{P}$s.

We conjecture that if we allow a discontinuous pricing function, $\hat{P}(P^e, G)$, then there exist Markov equilibria with $R > 1$, even under the conditions of this proposition.

4. MARKOV EQUILIBRIUM IN A LIMITED-PARTICIPATION MODEL

Here we analyze the set of Markov equilibria in a limited-participation model. We briefly describe the model. The sequence of events is as follows. Households start each period with nominal assets, and they must choose how much to deposit in a financial intermediary. The monetary authority then chooses its transfer to the financial intermediary. The financial intermediary makes loans to firms, who must borrow the wage bill before they produce. Households make their consumption and labor supply decision and firms make production decisions. Money is not neutral because households cannot change their deposit decision after the monetary authority chooses its transfer. Let Q denote the aggregate deposits made by households, and let G denote the growth rate of the money supply chosen by the monetary authority. In this section, as in the previous section, all prices and quantities of nominal assets are scaled by the aggregate stock of money. Let $S = (Q, G)$ denote the state of the economy after these decisions are made.

The household's utility function is

$$\sum_{t=0}^{\infty} \beta^t u(c_t, n_t), \quad u(c, n) = \log(c) + \gamma \log(1 - n),$$

where c_t and n_t denote date t consumption labor, respectively. We write the household's problem recursively. We start with the problem solved by the household after the monetary authority has made its transfer. Let A denote the household's beginning-of-period nominal assets. Let q denote its deposits. Both variables have been scaled by the aggregate, beginning-of-period stock of money. The consumption, employment, and asset accumulation decisions solve

$$w(A, q, S) = \max_{c,n,M'} u(c, n) + \beta v(A'),$$

subject to

$$P(S)c \leq W(S)n + A - q$$

and

$$GA' = R(S)[q + (G - 1)] - D(S) + W(S)n + A - q - P(S)c.$$

Here, v is the value function at the beginning of the next period, before the household makes next period's deposit decision. Also, $R(S)$ is the gross interest rate, $P(S)$ is the price of the consumption good, $W(S)$ is the wage rate, and

$D(S)$ is the profit from firms. The choice of q solves the following dynamic programming program:

$$v(A) = \max_q w(A, q, S^e),$$

where S^e is the state if the monetary authority does not deviate, that is, $S^e = [Q, X(Q)]$, where $X(Q)$ is the monetary authority's policy function.

The production sector is exactly as in the cash–credit good model, with one exception. To pay for the labor that they hire during the period, intermediate good producing firms must borrow in advance from the financial intermediary at gross interest rate $R(S)$. Thus, the marginal dollar cost of hiring a worker is $R(S)W(S)$, so that, by the type of reasoning in the cash–credit good model, we find $R(S)W(S)/P(S) = \lambda$.

The financial intermediary behaves competitively. It receives Q from households, and $G - 1$ on households' behalf from the central bank. When $R(S) > 1$, it lends all these funds in the loan market. When $R(S) = 1$, it supplies whatever is demanded, up to the funds it has available. We shall say that when $R(S) = 1$ and demand is less than available funds, then there is a "liquidity trap." At the end of the period the financial intermediary returns its earnings, $R(S)(Q + G - 1)$, to the households. Finally, if $R(S) < 1$, the financial intermediary lends no funds, and it returns $Q + G - 1$ to households. Loan demand by firms is given by $W(S)n(S)$. Therefore loan market clearing requires

$$W(S)n(S) \leq Q + G - 1,$$

with equality if $R(S) > 1$.

The monetary authority's policy function, $X(Q)$, solves

$$X(Q) \in \arg\max_G w(1, Q, Q, G).$$

A recursive private sector equilibrium and a Markov equilibrium are defined analogously to those in the previous section.

It is useful to begin with an analysis of outcomes under commitment. It is easy to show, as in Section 2, that the Ramsey equilibrium has $R = 1$ and can be supported by a policy that sets the growth rate of the money supply equal to β. Let c^*, n^*, W^*, R^*, P^*, and Q^* denote this Ramsey equilibrium. These variables solve the following system of equations:

$$\begin{aligned} &\frac{\gamma c^*}{1 - n^*} = \frac{W^*}{P^*}, \frac{W^*}{P^*} = \frac{\lambda}{R^*}, R^* = 1, \\ &W^* n^* = Q^* + \beta - 1, P^* c^* = W^* n^* + 1 - Q^*, \\ &c^* = n^*. \end{aligned} \tag{4.1}$$

It is straightforward to verify that the usual nonnegativity constraints are satisfied. Notice that the first equation is the household's first-order condition for labor, the second results from firm optimization, the third corresponds to the intertemporal Euler equation, the fourth corresponds to money market

clearing, the fifth is the household's cash-in-advance constraint, and the last equation corresponds to goods market clearing.

Next, we analyze the Markov equilibria of our model. The necessary and sufficient conditions for allocations and pricing functions to constitute a recursive private sector equilibrium are as follows:

$$\frac{\gamma n(S)}{1 - n(S)} = \frac{W(S)}{P(S)}, \tag{4.2}$$

$$\frac{W(S)}{P(S)} = \frac{\lambda}{R(S)}, \tag{4.3}$$

$$W(S)n(S) \leq \begin{cases} Q + G - 1 & \text{if } R(S) \geq 1 \\ 0 & \text{if } R(S) < 1 \end{cases}, \tag{4.4}$$

$$P(S)n(S) - W(S)n(S) \leq 1 - Q, \tag{4.5}$$

where (4.4) holds with equality if $R(S) > 1$. As already noted, if $R(S) < 1$, the supply of funds in the loan market is zero. Also, (4.5) holds with equality if $R[Q, X(Q)] > 1$ and $S = [Q, X(Q)]$. That is, if along the Markov equilibrium path the net interest rate is strictly positive, the household's cash-in-advance constraint is satisfied as a strict equality. In a deviation from the Markov equilibrium path, the cash-in-advance constraint must hold as a weak inequality, regardless of the realized interest rate.

We now establish the following proposition:

Proposition 4.1 (all Markov equilibria are Ramsey). *In any Markov equilibrium, $R[Q, X(Q)] = 1$, and the allocations and prices on the equilibrium path are the Ramsey outcomes given in (4.1).*

Proof. We prove this proposition in two parts. First, we construct a Markov equilibrium in which $R[Q, X(Q)] = 1$. Then, we show that there is no equilibrium with $R[Q, X(Q)] > 1$. Our constructed Markov equilibrium is as follows. Let $Q = Q^*$, where Q^* solves (4.1). On the equilibrium path, the monetary authority's decision rule is $X(Q^*) = \beta$. The allocation and pricing functions, $c(S)$, $n(S)$, $W(S)$, $P(S)$, and $R(S)$, in a recursive private sector equilibrium are defined as follows. For all S, $c(S) = n(S)$. For $G \leq \beta$, $R(S) = 1$, $n(S) = n^*$, $W(S)$ is obtained from (4.4) with equality, and $P(S) = W(S)/\lambda$. It is then easy to show that (4.5) holds with inequality. For $G > \beta$ the functions are defined as follows: $n(S) = n^*$, $W(S) = w^*$, $R(S) = R^* = 1$, and $P(S) = P^*$, where the variables with the asterisk are those associated with the Ramsey equilibrium, (4.1). Notice that these allocation and pricing rules satisfy (4.2), (4.3), and (4.5) with equality and (4.4) with inequality.

Next we show by contradiction that there does not exist a Markov equilibrium with $R[Q, X(Q)] > 1$. Suppose, to the contrary, that there did exist such an equilibrium. Notice that it is always possible to construct a private sector equilibrium for arbitrary $G \geq \beta$ by simply setting (4.2)–(4.5) to equality.

Therefore, the domain of deviation that has to be considered includes all $G > X(Q)$. Consider such a deviation. We will show that, in the private sector equilibrium associated with this deviation, $R(Q, G) < R[Q, X(Q)]$. This argument is also by contradiction. Thus, suppose $R(Q, G) \geq R[Q, X(Q)]$. Because $R[Q, X(Q)] > 1$, (4.4) must hold as an equality at the deviation. Substituting for $P(S)$ from (4.3) and $W(S)n(S)$ from (4.4), we see that the left side of (4.5) becomes

$$[R(S)/\lambda - 1](Q + G - 1),$$

which is larger than $[R(S)/\lambda - 1][Q + X(Q) - 1]$. On the equilibrium path, (4.5) must hold as an equality. Therefore, at the deviation (4.5) must be violated. We have established that, in any deviation of the form $G > X(Q)$, $R(Q, G) < R[Q, X(Q)]$. However, from (4.2) and (4.3), this raises employment toward the efficient level, contradicting monetary authority optimization. We have established the desired contradiction. ■

Notice that, in the Markov equilibrium we have constructed, there is a liquidity trap. If the monetary authority deviates and chooses a growth rate for the money supply greater than β, the resulting transfers of money are simply hoarded by the financial intermediary and not lent out to firms. All allocations and prices are unaffected by such a deviation.

5. CONCLUSION

In this paper we worked with an environment that, with one exception, is similar in spirit to the one analyzed in the Kydland–Prescott and Barro–Gordon literature. The exception is that we are explicit about the mechanisms that cause unanticipated monetary injections to generate benefits and distortions. We found that, for two standard models, there is no inflation bias at all.

ACKNOWLEDGMENTS

S. Albanesi is with Bocconi University IGIER, and CEPR; V. Chari is with the University of Minnesota and Federal Reserve Bank of Minneapolis; and L. Christiano is with Northwestern University and the Federal Reserve Banks of Chicago and Cleveland. Chari and Christiano thank the National Science Foundation for supporting this research.

APPENDIX A: NECESSARY AND SUFFICIENT CONDITIONS FOR HOUSEHOLD OPTIMIZATION IN THE CASH–CREDIT GOOD MODEL

This appendix develops necessary and sufficient conditions for optimality of the household problem in the cash–credit good model of Section 2. It is included

here for completeness. Many of the results here can be found in the literature. See, for example, Woodford (1994).

In what follows, we assume that

$$P_{1t}, P_{2t}, W_t > 0, R_t \geq 1, \lim_{t\to\infty} \sum_{j=0}^{t} q_{j+1}[W_j + T_j + D_j] \text{ finite.} \tag{A.1}$$

If these conditions did not hold, there could be no equilibrium. We begin by proving a proposition that allows us to rewrite the household's budget set in a more convenient form. We show the following.

Proposition A.1. *Suppose (2.2), (2.4), and (A.1) are satisfied. The constraint given in (2.5) is equivalent to*

$$\lim_{T\to\infty} q_T A_T \geq 0. \tag{A.2}$$

Proof. It is useful to introduce some new notation. Let I_t and S_t be defined by

$$I_t \equiv W_t + T_t + D_t \tag{A.3}$$

and

$$S_t \equiv (R_t - 1)M_t + P_{1t}c_{1t} + P_{2t}c_{2t} + W_t(1 - n_t),$$

respectively. It is straightforward to show that household nominal assets satisfy

$$A_{t+1} = I_t + R_t A_t - S_t. \tag{A.4}$$

We establish that (A.2) implies (2.5). Recursively solving for assets using (A.4) and (2.2) from t to T yields

$$q_T A_T \leq \sum_{j=0}^{T-t-1} q_{t+j-1} I_{t+j} + q_t A_t - \sum_{j=0}^{T-t-1} q_{t+j+1} S_{t+j}. \tag{A.5}$$

Taking into account $q_{t+j+1}S_{t+j} \geq 0$ and rewriting this expression, we obtain

$$q_t A_t \geq q_T A_T - \sum_{j=0}^{T-t-1} q_{t+j+1} I_{t+j}.$$

Fixing t, taking the limit, $T \to \infty$, and using (A.2) yields (2.5).

We now show that (2.5) implies (A.2). Note first that the limit in (A.1) being finite implies

$$\lim_{t\to\infty} \sum_{j=1}^{\infty} q_{t+j+1} I_{t+j} = 0.$$

Using this result and (2.5), we see that (A.2) follows trivially. ■

Following is the main result of this appendix.

Proposition A.2. *A sequence, $\{c_{1t}, c_{2t}, n_t, M_t, B_t\}$, solves the household problem if and only if the following conditions are satisfied. The Euler equations are*

$$\frac{u_{1t}}{P_{1t}} = R_t \frac{u_{2t}}{P_{2t}}, \tag{A.6}$$

$$\frac{-u_{3t}}{u_{2t}} = \frac{W_t}{P_{2t}}, \tag{A.7}$$

$$\frac{u_{1t}}{P_{1t}} = \beta R_t \frac{u_{1t+1}}{P_{1t+1}}, \tag{A.8}$$

$$(R_t - 1)(P_{1t}c_{1t} - M_t) = 0. \tag{A.9}$$

The transversality condition is (A.2) with equality:

$$\lim_{t\to\infty} q_t A_t = 0. \tag{A.10}$$

Proof. We begin by showing that if a sequence $\{c_{1t}, c_{2t}, n_t, M_t, B_t\}$ satisfies (A.6)–(A.10), then that sequence solves the household's problem. That is, we show that

$$D = \lim_{T\to\infty} \left[\sum_{t=0}^{T} \beta^t u(c_{1t}, c_{2t}, n_t) - \sum_{t=0}^{T} \beta^t u(c'_{1t}, c'_{2t}, n'_t) \right] \geq 0,$$

where $\{c'_{1t}, c'_{2t}, n'_t, M'_t, B'_t\}_{t=0}^{\infty}$ is any other feasible plan. Note first that the Euler equations imply

$$\beta^t u_{1,t} = q_t P_{1t} \frac{u_{1,0}}{P_{1,0}},$$

$$\beta^t u_{2,t} = q_{t+1} P_{2t} \frac{u_{1,0}}{P_{1,0}},$$

$$\beta^t u_{3,t} = -q_{t+1} W_t \frac{u_{1,0}}{P_{1,0}},$$

where $u_{i,t}$ is the derivative of u with respect to its ith argument. By concavity and the fact that the candidate optimal plan satisfies (A.6) and (A.7), we can write

$$\begin{aligned} D &\geq \lim_{T\to\infty} \frac{u_{1,0}}{P_{1,0}} \sum_{t=0}^{T} [q_t P_{1t}(c_{1t} - c'_{1t}) + q_{t+1} P_{2t}(c_{2t} - c'_{2t}) - q_{t+1} W_t (n_t - n'_t)] \\ &= \lim_{T\to\infty} \frac{u_{1,0}}{P_{1,0}} \sum_{t=0}^{T} q_t \left[\frac{S_t}{R_t} + \frac{1 - R_t}{R_t} (M_t - P_{1t} c_{1t}) \right. \\ &\quad \left. - \frac{S'_t}{R_t} - \frac{1 - R_t}{R_t} \left(M'_t - P_{1t} c'_{1t} \right) \right] \\ &\geq \lim_{T\to\infty} \frac{u_{1,0}}{P_{1,0}} \sum_{t=0}^{T} [q_{t+1} S_t - q_{t+1} S'_t], \end{aligned} \tag{A.11}$$

where the equality is obtained by using the definition of S_t and the second inequality is obtained by using $R_t \geq 1$, and $(1 - R_t)(M'_t - P_{1t}c'_{1t}) \leq 0$; see

(2.3). Iterating on (A.4) for the two plans, we can rewrite (A.11) as

$$\begin{aligned} D &\geq \lim_{T\to\infty} \frac{u_{1,0}}{P_{1,0}} \left[\sum_{t=0}^{T} q_{t+1} S_t + q_{T+1} A'_{T+1} - \sum_{t=0}^{T} q_{t+1} I_t - A_0 \right] \\ &\geq \lim_{T\to\infty} \frac{u_{1,0}}{P_{1,0}} \left[\sum_{t=0}^{T} q_{t+1} S_t - \sum_{t=0}^{T} q_{t+1} I_t - A_0 \right] \\ &= \lim_{T\to\infty} \frac{u_{1,0}}{P_{1,0}} q_{T+1} A_{T+1} \geq 0, \end{aligned}$$

by (A.10).

Now we establish that if $\{c_{1t}, c_{2t}, n_t, M_t, B_t\}$ is optimal, then (A.6)–(A.10) are true. That (A.6)–(A.9) are necessary is obvious. It remains to show that (A.10) is necessary. Suppose (A.10) is not true. We show this contradicts the hypothesis of optimality. We need only consider the case where $\lim_{T\to\infty} q_T A_T$ is strictly positive. The strictly negative case is ruled out by the preceding proposition. So, suppose

$$\lim_{T\to\infty} q_T A_T = \Delta > 0.$$

We construct a deviation from the optimal sequence that is consistent with the budget constraint and results in an increase in utility. Fix some particular date, τ. We replace $c_{1\tau}$ by $c_{1\tau} + \varepsilon/P_{1\tau}$, where $0 < \varepsilon \leq \Delta/q_\tau$. Consumption at all other dates and $c_{2\tau}$ are left unchanged, as well as employment at all dates. We finance this increase in consumption by replacing M_τ^d with $M_\tau^d + \varepsilon$ and B_τ with $B_\tau - \varepsilon$. Money holdings at all other dates are left unchanged. Debt and wealth after t, B_t, A_t, $t > \tau$, are different in the perturbed allocations. We denote the variables in the perturbed plan with a prime. From (A.4),

$$\begin{aligned} A'_{\tau+1} - A_{\tau+1} &= -R_\tau \varepsilon = -\frac{q_\tau}{q_{\tau+1}} \varepsilon, \\ A'_{\tau+j} - A_{\tau+j} &= -R_{\tau+j-1} \cdots R_\tau \varepsilon = -\frac{q_\tau}{q_{\tau+1}} \varepsilon. \end{aligned}$$

Multiplying this last expression by $q_{\tau+j}$ and setting $T = \tau + j$, we have

$$q_T(A'_T - A_T) = -q_\tau \varepsilon.$$

Taking the limit, as $T \to \infty$, we find

$$\lim_{T\to\infty} q_T A'_T = \Delta - q_\tau \varepsilon \geq 0.$$

We conclude that the perturbed plan satisfies (A.2). However, utility is clearly higher in the perturbed plan. We have a contradiction. ■

APPENDIX B: PROPERTIES OF MARKOV EQUILIBRIUM

In this appendix we prove Proposition 3.3. We establish the result by constructing a Markov equilibrium that supports the Ramsey outcomes. Specifically, we

construct P^e, a set of private sector allocation rules, a set of pricing functions, and a monetary policy rule, all of which satisfy the conditions for a Markov equilibrium. In Subsection 3.3 it is shown that private sector allocation rules and pricing functions can equivalently be expressed as functions of the growth rate of the money supply, G, or of $\hat{P}$, the price of the flexibly priced intermediate goods. Because these representations are equivalent and it is convenient to work with $\hat{P}$, we do so here.

The construction of the Markov equilibrium is as follows. Let c_1^*, c_2^*, W^*, R^*, P^*, P_1^*, and P_2^* solve (3.7)–(3.12) with $R = 1$ and with the cash-in-advance constraint holding with equality. That is, they are given by

$$c_1^* = \left[1 + \left(1 + \frac{\gamma}{\lambda}\right)\left(\frac{1-\alpha}{\alpha}\right)^{1/(1-\rho)} + \frac{\gamma}{\lambda}\right]^{-1}, \tag{B.1}$$

$$c_2^* = c_1^*\left(\frac{1-\alpha}{\alpha}\right)^{1/(1-\rho)}, \tag{B.2}$$

$R^* = 1$, $P_1^* = P_2^* = P^* = 1/c_1^*$, and $W^* = \lambda P^*$. Let $P^e = P^*$. For $\hat{P} > P^e$, let the allocation rules and pricing functions solve (3.7)–(3.12) with (3.8) replaced by $c_1 = 1/P_1$. For $\hat{P} < P^e$, let allocation rules and pricing functions solve (3.7)–(3.12) with $R = 1$. By construction, P^e and these allocation and pricing functions satisfy private sector optimality and market clearing. We need only check optimality of the monetary authority.

Denote the derivative of U in (3.15) with respect to $\hat{P}$ by L, where

$$L = u_1 c_1' + u_2 c_2' + u_n n', \tag{B.3}$$

where u_1, u_2, and u_n denote derivatives of the utility function with respect to the cash good, the credit good, and employment, respectively. In addition, c_1', c_2', and n' denote derivatives of the allocation rules defined in (3.14) with respect to $\hat{P}$. These derivatives and all others in this appendix are evaluated at $\hat{P} = P^e$. Let L^+ be the right derivative and L^- be the left derivative associated with L. We show that when our sufficient conditions are met, $L^+ \leq 0$ and $L^- \geq 0$.

Note that

$$P_i' = (1 - \mu_i), \quad i = 1, 2. \tag{B.4}$$

Using (B.4) and grouping terms in (B.3), we obtain

$$\begin{aligned} L &= u_2\left[\frac{u_1}{u_2}c_1' + c_2' + \frac{u_3}{u_2}(c_1' + c_2')\right] \\ &= (1-\lambda)u_2 c_1\left[\frac{c_1'}{c_1} + \frac{c_2}{c_1}\frac{c_2'}{c_2}\right], \end{aligned} \tag{B.5}$$

because $u_1/u_2 = R = 1$ and $-u_3/u_2 = \lambda$, when $\hat{P} = P^e$.

B.1. Right Derivative

We now establish that when our sufficient conditions are met, $L^+ \leq 0$. To evaluate the derivatives in (B.5), we require expressions for c_1'/c_1 and c_2'/c_2. The first of these is obtained by differentiating the binding cash-in-advance constraint:

$$\frac{c_1'}{c_1} = -\frac{1-\mu_1}{P^e}. \tag{B.6}$$

To obtain c_2'/c_2, note that the static labor Euler equation is given by

$$\frac{\gamma c_2}{1-n}\left(\frac{c}{c_2}\right)^{\rho} = \lambda\frac{\hat{P}}{P_2},$$

or, substituting for c and rearranging,

$$\frac{\gamma}{1-\alpha}\left[\alpha\left(\frac{c_1}{c_2}\right)^{\rho} + 1 - \alpha\right] = \lambda\frac{\hat{P}}{P_2}\frac{1-n_1-n_2}{c_2}. \tag{B.7}$$

Differentiating both sides of this expression with respect to $\hat{P}$ and taking into account $d(\hat{P}/P_2)/d\hat{P} = \mu_2/P_2$ when $\hat{P} = P^e$, we obtain, after some manipulations,

$$\left[\lambda\frac{1-c_1}{c_1} - \gamma\rho\right]\frac{c_2'}{c_2} = \lambda\frac{\mu_2}{P_2}\frac{1-c_1-c_2}{c_1} - [\lambda+\gamma\rho]\frac{c_1'}{c_1}. \tag{B.8}$$

Substituting for c_1'/c_1 and c_2'/c_2 from (B.6) and (B.8), respectively, into (B.5), we obtain

$$\begin{aligned} L^+ = \frac{(1-\lambda)u_2c_1}{\lambda[(1-c_1)/c_1] - \gamma\rho}\Bigg\{&\left[\frac{c_2}{c_1}(\lambda+\gamma\rho) - \left(\lambda\frac{1-c_1}{c_1} - \gamma\rho\right)\right] \\ &\times\frac{1-\mu_1}{P^e} + \frac{c_2}{c_1}\left(\lambda\frac{\mu_2}{P_2}\frac{1-c_1-c_2}{c_1}\right)\Bigg\}. \end{aligned} \tag{B.9}$$

The denominator of (B.9) is positive. To see this, use (B.1) to show

$$\lambda\frac{1-c_1}{c_1} - \gamma\rho = \lambda\left(1+\frac{\gamma}{\lambda}\right)\left(\frac{1-\alpha}{\alpha}\right)^{1/(1-\rho)} + \gamma(1-\rho) > 0, \tag{B.10}$$

because $\rho \leq 1$. We can rewrite (B.9) as

$$\begin{aligned} L^+ = \frac{u_2c_2(1-\lambda)}{P_2\{\lambda[(1-c_1)/c_1] - \gamma\rho(g/\beta)\}}\Bigg\{&-\left[\lambda\frac{1-c_1-c_2}{c_2} - \gamma\rho\frac{c_1+c_2}{c_2}\right] \\ &\times(1-\mu_1) + \lambda\mu_2\frac{1-c_1-c_2}{c_1}\Bigg\}. \end{aligned} \tag{B.11}$$

Substituting for c_1 from (B.1) and c_2/c_1 from (B.2), we obtain

$$\frac{1-c_1-c_2}{c_1} = \frac{\gamma}{\lambda}\left[\left(\frac{1-\alpha}{\alpha}\right)^{1/(1-\rho)} + 1\right]. \tag{B.12}$$

In addition,

$$\frac{1-c_1-c_2}{c_2} = \frac{\gamma}{\lambda}\left[1 + \left(\frac{1-\alpha}{\alpha}\right)^{-1/(1-\rho)}\right],$$

and

$$\frac{c_1+c_2}{c_2} = \left(\frac{1-\alpha}{\alpha}\right)^{-1/(1-\rho)} + 1.$$

Substituting these results into (B.11), we obtain

$$\begin{aligned} L^+ = {} & \frac{u_2 c_2(1-\lambda)}{P_2\left\{\lambda[(1-c_1)/c_1] - \gamma\rho(g/\beta)\right\}} \\ & \times \left[-\left(\lambda\left\{\frac{\gamma}{\lambda}\left[1 + \left(\frac{1-\alpha}{\alpha}\right)^{-1/(1-\rho)}\right]\right\}\right.\right. \\ & \left. - \gamma\rho\left[\left(\frac{1-\alpha}{\alpha}\right)^{-1/(1-\rho)} + 1\right]\right)(1-\mu_1) \\ & \left. + \lambda\mu_2\left\{\frac{\gamma}{\lambda}\left[\left(\frac{1-\alpha}{\alpha}\right)^{1/(1-\rho)} + 1\right]\right\}\right]. \end{aligned}$$

Simplifying, we have

$$\begin{aligned} L^+ = {} & \frac{\gamma u_2 c_2(1-\lambda)\left\{1 + [(1-\alpha)/\alpha]^{-1/(1-\rho)}\right\}}{P_2\left\{\lambda[(1-c_1)/c_1] - \gamma\rho(g/\beta)\right\}} \\ & \times \left\{-(1-\rho)(1-\mu_1) + \mu_2\left(\frac{1-\alpha}{\alpha}\right)^{1/(1-\rho)}\right\}. \end{aligned}$$

Because the term in front of the large braces is positive, it follows that $L^+ \leq 0$ if and only if

$$(1-\rho)(1-\mu_1) \geq \mu_2\left(\frac{1-\alpha}{\alpha}\right)^{1/(1-\rho)}. \tag{B.13}$$

B.2. Left Derivative

Next, we establish that under the sufficient conditions of the proposition, $L^- \geq 0$. The expression for c_2'/c_2 is still given by (B.8). To obtain c_1'/c_1, we

differentiate (3.9) with $R = 1$ and use (B.4) to obtain

$$\frac{\alpha}{1-\alpha}(1-\rho)\left(\frac{c_2}{c_1}\right)^{1-\rho}\left(\frac{c_2'}{c_2}-\frac{c_1'}{c_1}\right)=\frac{\mu_2-\mu_1}{P}, \tag{B.14}$$

or, because $[\alpha/(1-\alpha)](c_2/c_1)^{1-\rho}=1$,

$$\frac{c_2'}{c_2}-\frac{c_1'}{c_1}=\frac{\mu_2-\mu_1}{(1-\rho)P}.$$

Substituting for c_1'/c_1 from here into (B.8) and collecting terms, we obtain, after simplifying,

$$\lambda\frac{1}{c_1}\frac{c_2'}{c_2}=\lambda\frac{\mu_2}{P_2}\frac{1-c_1-c_2}{c_1}+(\lambda+\gamma\rho)\frac{\mu_2-\mu_1}{(1-\rho)P}.$$

Then, using (B.12), we obtain

$$\lambda\frac{1}{c_1}\frac{c_2'}{c_2}=\frac{\mu_2}{P_2}\gamma\left[\left(\frac{1-\alpha}{\alpha}\right)^{1/(1-\rho)}+1\right]+(\lambda+\gamma\rho)\frac{\mu_2-\mu_1}{(1-\rho)P}.$$

Now substitute out for c_1'/c_1 and c_2'/c_2 into (B.5) to obtain, after simplifying,

$$L^{-}=(1-\lambda)\frac{u_2c_1\gamma}{P^e(\lambda+\gamma)}\left[\mu_2\left(\frac{1-\alpha}{\alpha}\right)^{1/(1-\rho)}+\mu_1\right]>0.$$

APPENDIX C: UNIQUENESS RESULT

We prove Proposition 3.4 by contradiction. Suppose that there exists a Markov equilibrium outcome with $R > 1$. The contradiction is achieved in two steps. First, we establish that a deviation down in $\hat{P}$ can be accomplished by some feasible deviation in G. We then establish that such a deviation is desirable. That a Markov equilibrium exists follows from Proposition 3.3.

C.1. Feasibility of a Downward Deviation in $\hat{P}$

Let P^e denote the expected price level in the Markov equilibrium, and let G^e denote the money growth rate in the corresponding equilibrium outcome, that is, $G^e = X(P^e)$. We establish that for any $\hat{P}$ in a neighborhood, U, of P^e, there exists a G belonging to a neighborhood, V, of G^e, such that $\hat{P} = \hat{P}(P^e, G)$. Here, $\hat{P}(P^e, G)$ is the price allocation rule in the Markov equilibrium.

Substituting from (3.9) into (3.13) and using the assumptions, $\sigma = 1$ and $\rho = 0$, we obtain

$$G(P^e, \hat{P}) = P_2(P^e, \hat{P})c_2(P^e, \hat{P})\frac{\beta}{1-\alpha}v_1[1, P^e, X(P^e)]. \tag{C.1}$$

From the analogs of (B.6) and (B.8) obtained for the case $g/\beta \geq 1$ and using $\rho = 0$, we can determine that $c_2(P^e, \hat{P})$ is a strictly increasing function of $\hat{P}$ for

$\hat{P}$ in a sufficiently small neighborhood, U, of P^e. It is evident from (3.12) that P_2 is globally increasing in $\hat{P}$. This establishes that $G(P^e, \hat{P})$ is strictly increasing for $\hat{P} \in U$. By the inversion theorem, $G(P^e, \hat{P})$ has a unique, continuous inverse function mapping from $V = G(P^e, U)$ to U. By continuity of $\hat{P}(P^e, G)$, this inverse is $\hat{P}(P^e, G)$ itself. This establishes the desired result.

C.2. Desirability of a Downward Deviation in $\hat{P}$

To show that a deviation, $\hat{P} < P^e$, is desirable, we first establish properties of the private sector allocation rules and pricing functions in Markov equilibria in which the interest rate is strictly greater than one. Let

$$x^a(P^e, \hat{P}) \equiv \left[c_1^a(P^e, \hat{P}), c_2^a(P^e, \hat{P}), \ldots, R^a(P^e, \hat{P})\right]$$

denote the solutions to (3.7)–(3.12) with (3.8) replaced by the cash-in-advance constraint holding with equality. Let

$$x^b(P^e, \hat{P}) \equiv \left[c_1^b(P^e, \hat{P}), c_2^b(P^e, \hat{P}), \ldots, R^b(P^e, \hat{P})\right]$$

denote the solutions to (3.7)–(3.12) with (3.8) replaced by $R = 1$. For any $\hat{P}$, P^e, private sector allocations and prices must be given either by $x^a(P^e, \hat{P})$ or $x^b(P^e, \hat{P})$. We now show that for all $\hat{P}$ in a neighborhood of P^e, the private sector allocations and prices must be given by $x^a(P^e, \hat{P})$.

Consider $\hat{P} = P^e$. Solving (3.7)–(3.12) with (3.8) replaced by the cash-in-advance constraint holding with equality and with $\hat{P} = P^e$, we obtain

$$c_1^a(P^e, P^e) = \left[1 + \left(1 + \frac{\gamma}{\lambda}\right)\left(\frac{1-\alpha}{\alpha}R\right)^{1/(1-\rho)} + \frac{\gamma}{\lambda}\right]^{-1}.$$

Solving the analogous equations for $c_1^b(P^e, P^e)$, we obtain

$$c_1^b(P^e, P^e) = \left[1 + \left(1 + \frac{\gamma}{\lambda}\right)\left(\frac{1-\alpha}{\alpha}\right)^{1/(1-\rho)} + \frac{\gamma}{\lambda}\right]^{-1}.$$

Evidently, $P^e c_1^b(P^e, P^e) > P^e c_1^a(P^e, P^e)$. By continuity, for all $\hat{P}$ in some neighborhood of P^e, we see $\hat{P}c_1^b(\hat{P}, P^e) > \hat{P}c_1^a(\hat{P}, P^e)$. Because $\hat{P}c_1^a(\hat{P}, P^e) = 1$, it follows that $\hat{P}c_1^b(\hat{P}, P^e) > 1$ for all $\hat{P}$ in a neighborhood of P^e. Because the cash-in-advance constraint is violated, $x^b(\hat{P}, P^e)$ cannot be part of a Markov equilibrium. We have established that for $\hat{P}$ in a neighborhood of P^e, private sector allocation rules and prices must be given by $x^a(\hat{P}, P^e)$.

With these pricing and allocation functions, the derivative of the utility function with respect to $\hat{P}$, evaluated at $P^e = \hat{P}$, can be shown to be

$$L = \frac{u_2 c_2 / P_2}{\lambda[(1-c_1)/c_1] - \gamma\rho\frac{g}{\beta}}\left[-a\left(\frac{g}{\beta}\right) + b\left(\frac{g}{\beta}\right)\right]$$

where g is the growth rate of money at the supposed outcome and

$$a\left(\frac{g}{\beta}\right) = \left(\frac{g}{\beta} - \lambda\right) \times \left[\lambda + \gamma + \gamma\left(\frac{g}{\beta}\right)^{[-\rho/(1-\rho)]}\left(\frac{1-\alpha}{\alpha}\right)^{[-1/(1-\rho)]}(1-\rho)\right](1-\mu_1),$$

$$b\left(\frac{g}{\beta}\right) = (1-\lambda)\gamma\mu_2\left[\left(\frac{g}{\beta}\frac{1-\alpha}{\alpha}\right)^{[1/(1-\rho)]} + \frac{g}{\beta}\right] + (1-\mu_1)(1-\lambda)\left(\lambda + \gamma\rho\frac{g}{\beta}\right).$$

Condition (ii) guarantees that $a(1) \geq b(1)$.[5] In addition under (i) a and b are linear with slopes a' and b', respectively, given by

$$a' = \left(\lambda + \frac{\gamma\alpha}{1-\alpha}\right)(1-\mu_1),$$

$$b' = \frac{(1-\lambda)\gamma\mu_2}{\alpha}.$$

Given (iii), it is trivial to verify that $L < 0$ for all $g/\beta > 1$. Thus, the supposition that there is an outcome with $R > 1$ leads to the implication that the monetary authority can raise utility by reducing $\hat{P}$. This contradicts monetary authority maximization. We conclude that there are no Markov equilibrium outcomes with $R > 1$.

References

Albanesi, S., V. V. Chari, and L. J. Christiano (2002), "Expectation Traps and Monetary Policy," NBER, working paper 8912.

Barro, R. J. and D. B. Gordon (1983), "A Positive Theory of Monetary Policy in a Natural Rate Model," *Journal of Political Economy*, 91, 589–610.

Blanchard, O. J. and N. Kiyotaki (1987), "Monopolistic Competition and the Effects of Aggregate Demand," *American Economic Review*, 77(4), 647–666.

Chari, V. V., L. J. Christiano, and M. Eichenbaum (1998), "Expectation Traps and Discretion," *Journal of Economic Theory*, 81(2), 462–492.

Christiano, L. J., M. Eichenbaum, and C. Evans (1997), "Sticky Price and Limited Participation Models of Money: A Comparison," *European Economic Review*, 41, 1201–1249.

Cole, H. and N. Kocherlakota (1998), "Zero Nominal Interest Rates: Why They Are Good and How to Get Them," *Federal Reserve Bank of Minneapolis Quarterly Review*, 22(2), 2–10.

[5] It is easily verified that this is equivalent to condition (B.13).

Ireland, P. (1997), "Sustainable Monetary Policies," *Journal of Economic Dynamics and Control*, 22, 87–108.

Kydland, F. and E. C. Prescott (1977), "Rules Rather than Discretion: The Inconsistency of Optimal Plans," *Journal of Political Economy*, 85(3), 473–491.

Lucas, R. E., Jr. and N. Stokey (1983), "Optimal Fiscal and Monetary Policy in an Economy without Capital," *Journal of Monetary Economics*, 12, 55–93.

Neiss, K. (1999), "Discretionary Inflation in a General Equilibrium Model," *Journal of Money, Credit, and Banking*, 31(3), 357–374.

Nicolini, J. P. (1998), "More on the Time Consistency of Monetary Policy," *Journal of Monetary Economics*, 41(2), 333–350.

Svensson, L. (1985), "Money and Asset Prices in a Cash-in-Advance Economy," *Journal of Political Economy*, 93(5), 919–944.

Woodford, M. (1994), "Monetary Policy and Price Level Determinacy in a Cash-in-Advance Economy," *Economic Theory*, 4, 345–380.

CHAPTER 5

New Perspectives on Monetary Policy, Inflation, and the Business Cycle

Jordi Galí

1. INTRODUCTION

In recent years, the field of macroeconomics has witnessed the development of a new generation of small-scale monetary business cycle models, generally referred to as New Keynesian (NK) models or New Neoclassical Synthesis models. The new models integrate Keynesian elements (imperfect competition, and nominal rigidities) into a dynamic general equilibrium framework that until recently was largely associated with the Real Business Cycle (RBC) paradigm. They can be used (and are being used) to analyze the connection among money, inflation, and the business cycle, and to assess the desirability of alternative monetary policies.

In contrast with earlier models in the Keynesian tradition, the new paradigm has adopted a dynamic general equilibrium modeling approach. Thus, equilibrium conditions for aggregate variables are derived from optimal individual behavior on the part of consumers and firms, and are consistent with the simultaneous clearing of all markets. From that viewpoint, the new models have much stronger theoretical foundations than traditional Keynesian models.

In addition, the emphasis given to nominal rigidities as a source of monetary nonneutralities also provides a clear differentiation between NK models and classical monetary frameworks.[1] In the latter, the key mechanism through which money may have some real effect is the so-called inflation tax. However, those effects are generally acknowledged to be quantitatively small and not to capture the main sources of monetary nonneutralities at work in actual economies.[2]

The purpose of the present paper is twofold. First, it tries to provide an overview of some of the recent developments in the literature on monetary policy in the presence of nominal rigidities. Given the voluminous literature

[1] By classical I mean monetary models with perfect competition and flexible prices, and no other frictions (other than those associated with the existence of money).

[2] See, for example, Cooley and Hansen (1989) for an analysis of a classical monetary model. Several authors have also emphasized the existence of frictions in financial markets as a potential source of nontrivial (and more realistic) monetary nonneutralities. See, for example, Christiano, Eichenbaum, and Evans (1997, 1998).

generated by researchers working in this area, and the usual space constraints, that overview will necessarily be partial.[3] Second, the paper seeks to emphasize the existence of several dimensions in which the recent literature provides a *new perspective* on the linkages among monetary policy, inflation, and the business cycle. I would like to argue that the adoption of an explicitly optimizing, general equilibrium framework has not been superfluous; on the contrary, the recent literature has yielded many genuinely new insights that, by their nature, could hardly have been obtained with, say, a textbook IS-LM model. Hence, and contrary to what some macroeconomists may believe, it is not all déja vu and we are not back to where we stood before Kydland and Prescott (1982).

For concreteness, let me summarize next some of the findings, ideas, or features of the new models that one may view as novel, relative to the traditional Keynesian literature. Needless to say, the list is not meant to be exhaustive; instead it focuses on some of the issues that are discussed in more detail in the remainder of the paper.

First, the NK models bring a new perspective on the nature of inflation dynamics. First, they emphasize the *forward-looking nature of inflation*. As argued in the paragraphs that follow, that property must be inherent to any model where prices are set by firms facing constraints on the frequency with which they can adjust the price of the goods they produce. Firms that are resetting their prices today recognize that the prices they choose are likely to remain effective for more than one period. Such firms will find it optimal, when making their current pricing decisions, to take into account their expectations regarding future cost and demand conditions. Because changes in the aggregate price level are (by definition) a consequence of current pricing decisions, it follows that inflation must have an important forward-looking component. That property appears clearly reflected in the so-called New Phillips Curve. As discussed later, it also appears to be a feature of the data. Second, NK models also stress the important role played by *variations in markups* (or, equivalently, in real marginal costs) as a source of changes in aggregate inflation. The latter can thus be interpreted as the consequence of firms' periodic attempts to correct the misalignment between actual and desired markups.

Second, the concept of *output gap* plays a central role in the new optimizing sticky price models, both as a force underlying fluctuations in inflation (through its influence on marginal costs) and as a policy target. However, the notion of output gap found in the recent literature bears little resemblance to the ad hoc,

[3] It is also likely to be somewhat biased in that it attaches a disproportionate weight to issues or areas in which I happen to have done some research myself. Alternative surveys of some of those developments, with a somewhat different focus, can be found in Goodfriend and King (1997) and Clarida, Galí, and Gertler (1999). The present paper does not discuss any of the open economy extensions of NK models, and the issues that openness brings about; the interested reader will find in Lane (2001) a useful survey of developments on that front.

largely atheoretical output-gap measures used in traditional empirical analyses of inflation and monetary policy. In the new paradigm, the output gap has a precise meaning: it is the deviation of output from its equilibrium level in the absence of nominal rigidities. Under some assumptions on technology and preferences, it is possible to construct a measure of the output gap. As shown in what follows, the resulting measure for the postwar United States shows little resemblance to traditional output-gap measures.

Third, in the NK model, the transmission of monetary policy shocks to real variables works through a conventional interest rate channel. Yet, such a transmission mechanism does not necessarily involve a *liquidity effect*, in contrast with the textbook IS-LM model.

Fourth, in addition to being a source of monetary nonneutralities, the presence of sticky prices may also have strong implications for the economy's response to *nonmonetary shocks*. In particular, recent research has shown that unless monetary policy is sufficiently accommodating, employment is likely to drop in the short run in response to a favorable technology shock. That result is at odds with the predictions of standard RBC models, and it contrasts sharply with the mechanisms underlying fluctuations emphasized in the RBC literature. Most interestingly, the prediction of a negative short-run comovement between technology and employment appears to be supported by some recent estimates of the effects of identified technology shocks.

Fifth, the adoption of a general equilibrium framework by the recent sticky price literature permits an explicit *utility-based welfare analysis* of the consequences of alternative monetary policies, and can thus be used as the basis for the design of an optimal (or, at least, desirable) monetary policy. Hence, in the baseline sticky price model developed in the paragraphs that follow, the *optimal policy stabilizes the price level and the output gap completely*. Such a goal is fully attainable, because the central bank does not face a trade-off between output gap and inflation stabilization. Interestingly, the optimality of a zero inflation arises independently from any desire to reduce the distortion associated with the so-called inflation tax; instead, it is exclusively motivated by the policy maker's attempt to offset the distortions associated with staggered price setting, in order to replicate the flexible price equilibrium allocation.

Sixth, although the optimal monetary policy requires that the central bank respond systematically to the underlying disturbances in a specific way, a *simple policy rule* that has the central bank adjust (sufficiently) the interest rate in response to variations in inflation and/or the output gap generally provides a good approximation to the optimal rule (with the implied welfare losses being small). This is generally not the case for other well-known simple rules, such as a constant money growth or an interest rate peg.

Seventh, an interesting new insight found in the recent literature relates to the issue of rules versus discretion, and the role of credibility in monetary policy. The main result can be summarized as follows: In the presence of a trade-off

between output and inflation, society will generally gain from having a central bank that can (credibly) commit to a state-contingent plan. Most interestingly, such *gains from commitment arise even in the absence of a classic inflation bias*, that is, even if the central bank has no desire to push output above its natural level. That result overturns an implication of the classic Barro–Gordon analysis, in which the gains from commitment arise only if the central bank sets a target for output that does not correspond to its natural level.

Eighth, and finally, the coexistence of *staggered wage setting* with staggered price setting has important implications for monetary policy. In particular, the variations in wage markups caused by wages that are not fully flexible generate a trade-off between output gap and inflation stabilization that is absent from the basic sticky price model. Furthermore, recent research has shown that, in such an environment, a central bank will generally be unable to completely eliminate the distortions caused by nominal rigidities. The optimal policy will seek to strike a balance between stabilization of three variables: the output gap, price inflation, and wage inflation.

The remainder of the paper is organized as follows. Section 2 lays out a baseline sticky price model and derives the corresponding equilibrium conditions. Section 3 focuses on one of the building blocks of that model, the New Phillips Curve, and discusses some of its implications and empirical relevance. Section 4 uses the baseline model to analyze the effects and transmission of monetary and technology shocks in the presence of sticky prices. Section 5 turns its attention to the endogenous component of monetary policy: It derives the optimal policy rule and assesses the implications of deviating from it by following a simple rule instead. It also shows how the form of the optimal policy is altered by the presence of an inflation–output trade-off, and it discusses the gains from commitment that arise in that context. Section 6 brings sticky wages into the picture and analyzes consequences for the effects of monetary policy and its optimal design.

2. MONEY AND STICKY PRICES: A BASELINE MODEL

In this section I lay out a simple model that I take as representative of the new generation of dynamic sticky price models. It is a version of the Calvo (1983) model with staggered price setting.[4] For simplicity, and to focus on the essential aspects of the model, I work with a simplified version that abstracts

[4] Alternative approaches to modeling price rigidities have been used in the literature. Those include (a) models with staggered price setting à la Taylor (with a certain time between price adjustments), as exemplified by the work of Chari, Kehoe, and McGrattan (1996), and (b) models with convex costs of price adjustment (but no staggering), as in Hairault and Portier (1993) and Rotemberg (1996).

from capital accumulation and the external sector. Next I briefly describe the main assumptions, and I derive the key equilibrium conditions.[5]

2.1. Households

The representative consumer is infinitely lived and seeks to maximize

$$E_0 \sum_{t=0}^{\infty} \beta^t \left(\frac{C_t^{1-\sigma}}{1-\sigma} - \frac{N_t^{1+\varphi}}{1+\varphi} \right) \tag{2.1}$$

subject to a (standard) sequence of budget constraints and a solvency condition. N_t denotes hours of work. C_t is a constant elasticity of substitution (CES) aggregator of the quantities of the different goods consumed:

$$C_t = \left[\int_0^1 C_t(i)^{(\varepsilon-1)/\varepsilon} di \right]^{\varepsilon/(\varepsilon-1)}.$$

Let $P_t = [\int_0^1 P_t(i)^{1-\varepsilon}\, di]^{1/(1-\varepsilon)}$ represent the aggregate price index, where $P_t(i)$ denotes the price of good $i \in [0, 1]$. The solution to the consumer's problem can be summarized by means of three optimality conditions (two static and one intertemporal), which I represent in log-linearized form (henceforth, lowercase letters denote the logarithm of the original variables).

First, the optimal allocation of a given amount of expenditures among the different goods generated by the set of demand schedules implies

$$c_t(i) = -\varepsilon(p_t(i) - p_t) + c_t. \tag{2.2}$$

Second, and under the assumption of a perfectly competitive labor market, the supply of hours must satisfy

$$w_t - p_t = \sigma c_t + \varphi n_t, \tag{2.3}$$

where w is the (log) nominal wage.

Finally, the intertemporal optimality condition is given by the Euler equation:

$$c_t = -\frac{1}{\sigma}(r_t - E_t\{\pi_{t+1}\} - \rho) + E_t\{c_{t+1}\}, \tag{2.4}$$

where r_t is the yield on a nominally riskless one-period bond (the nominal interest rate, for short), π_{t+1} is the rate of inflation between t and $t+1$, and $\rho = -\log \beta$ represents the time discount rate (as well as the steady-state real interest rate, given the absence of secular growth).

Let me also postulate, without deriving it, a standard money demand equation:

$$m_t - p_t = y_t - \eta r_t, \tag{2.5}$$

[5] See, e.g., King and Wolman (1996), Yun (1996), and Woodford (1996) for a detailed derivation of the model's equilibrium conditions.

which will be used in some of the exercises described in what follows. Notice that a unit income elasticity of money demand is assumed, which is in line with much of the existing empirical evidence.

2.2. Firms

I assume a continuum of firms, each producing a differentiated good with a technology

$$Y_t(i) = A_t N_t(i),$$

where (log) productivity $a_t = \log(A_t)$ follows an exogenous, difference-stationary stochastic process represented by

$$\Delta a_t = \rho_a \Delta a_{t-1} + \varepsilon_t^a,$$

where $\{\varepsilon_t^a\}$ is white noise and $\rho_a \in [0, 1)$.

I assume that employment is subsidized at a constant subsidy rate ν. Hence, all firms face a common *real* marginal cost, which in equilibrium is given by

$$mc_t = w_t - p_t - a_t - \nu. \tag{2.6}$$

Total demand for each good is given by

$$Y_t(i) = C_t(i) + G_t(i),$$

where G_t denotes government purchases. For simplicity, I assume that the government consumes a fraction τ_t of the output of each good. Hence, and letting $g_t = -\log(1 - \tau_t)$, we can rewrite the demand for good i in log form as follows[6]:

$$y_t(i) = c_t(i) + g_t.$$

Let $Y_t = (\int_0^1 Y_t(i)^{(\varepsilon-1)/\varepsilon} di)^{\varepsilon/(\varepsilon-1)}$ denote aggregate output. The clearing of all goods markets implies

$$y_t = c_t + g_t, \tag{2.7}$$

where $y_t = \log Y_t$. In what follows I assume a simple AR(1) process for the demand shock g_t,

$$g_t = \rho_g g_{t-1} + \varepsilon_t^g,$$

where $\{\varepsilon_t^g\}$ is white noise (and orthogonal to ε_t^a) and $\rho_g \in [0, 1)$.

Euler equation (2.4), combined with market clearing, yields the equilibrium condition

$$y_t = -\frac{1}{\sigma}(r_t - E_t\{\pi_{t+1}\} - \rho) + E_t\{y_{t+1}\} + (1 - \rho_g)\, g_t. \tag{2.8}$$

[6] One can also interpret g_t as a shock to preferences or, more broadly, as an exogenous component of aggregate demand.

In addition, and using the fact that $n_t = \log \int_0^1 N_t(i)di$, we can derive the following mapping between labor input and output aggregates[7]:

$$n_t = y_t - a_t. \tag{2.9}$$

Finally, combining (2.3), (2.6), (2.7), and (2.9), we obtain an expression for the equilibrium real marginal cost in terms of aggregate output and productivity:

$$mc_t = (\sigma + \varphi)y_t - (1 + \varphi)a_t - \sigma g_t - \nu. \tag{2.10}$$

Notice that in deriving all of these equilibrium relationships, I have not made use of any condition specifying how firms set their prices. Next I describe two alternative models of price setting, which differ in the existence or not of restrictions on the frequency with which firms may adjust prices.

2.3. Flexible Price Equilibrium

Suppose that all firms adjust prices optimally each period, taking the path of aggregate variables as given. The assumption of an isoelastic demand implies that they will choose a markup (defined as the ratio of price to marginal cost) given by $\varepsilon/(\varepsilon - 1)$. That markup will be common across firms and constant over time. Hence, it follows that the real marginal cost (i.e., the inverse of the markup) will also be constant, and given by

$$mc_t = -\mu$$

for all t, where $\mu = \log[\varepsilon/(\varepsilon - 1)]$.[8] Furthermore, given identical prices and demand conditions, the same quantities of all goods will be produced and consumed.

In that case the equilibrium processes for output, consumption, hours, and the expected real rate are given by

$$\bar{y}_t = \gamma + \psi_a a_t + \psi_g g_t, \tag{2.11}$$

$$\bar{c}_t = \gamma + \psi_a a_t - (1 - \psi_g)g_t, \tag{2.12}$$

$$\bar{n}_t = \gamma + (\psi_a - 1)a_t + \psi_g g_t, \tag{2.13}$$

$$\overline{rr}_t = \rho + \sigma\psi_a\rho_a\Delta a_t + \sigma(1 - \psi_g)(1 - \rho_g)g_t, \tag{2.14}$$

where $\psi_a = (1 + \varphi)/(\sigma + \varphi)$, $\psi_g = \sigma/(\sigma + \varphi)$, and $\gamma = (\nu - \mu)/(\sigma + \varphi)$. Henceforth, I refer to these equilibrium values as the *natural* levels of the corresponding variable.

Notice that, in the absence of nominal rigidities, the equilibrium behavior of these variables is independent of monetary policy. Furthermore, if $\gamma = 0$,

[7] For nondegenerate distributions of prices across firms, the previous equation holds only up to a first-order approximation. More generally, we have $n_t = y_t - a_t + \xi_t$, where $\xi_t \equiv \log \int_0^1 \{[P_t(i)]/(P_t)\}^{-\varepsilon} di$ can be interpreted as an indicator of relative price distortions. See Yun (1996) and King and Wolman (1996) for a detailed discussion.

[8] Henceforth, an overbar is used to denote the equilibrium value of a variable under flexible prices.

the equilibrium allocation under flexible prices coincides with the *efficient* allocation, that is, the one that would be obtained under flexible prices, perfect competition, and no distortionary taxation of employment (i.e., $\nu = \mu = 0$). Attaining that efficient allocation requires setting $\nu = \mu$, that is, using an employment subsidy that exactly offsets the distortion associated with monopolistic competition. As discussed further in what follows, that assumption will generally be maintained.[9]

2.4. Staggered Price Setting

The exact form of the equation describing aggregate inflation dynamics depends on the way sticky prices are modeled. Let me follow Calvo (1983), and assume that each firm resets its price in any given period only with probability $1 - \theta$, independent of other firms and of the time elapsed since the last adjustment. Thus, a measure $1 - \theta$ of producers reset their prices each period, while a fraction θ keep their prices unchanged. Let p_t^* denote the log of the price set by firms adjusting prices in period t.[10] The evolution of the price level over time can be approximated by the log-linear difference equation:

$$p_t = \theta p_{t-1} + (1 - \theta) p_t^*. \tag{2.15}$$

One can show that a firm seeking to maximize its value will choose the price of its good according to the (approximate) log-linear rule

$$p_t^* = \mu + (1 - \beta\theta) \sum_{k=0}^{\infty} (\beta\theta)^k E_t\{\mathrm{mc}_{t+k}^n\}, \tag{2.16}$$

that is, prices are set as a markup over a weighted average of current and expected future nominal marginal costs $\{\mathrm{mc}_{t+k}^n\}$.

To get some intuition for the form of that rule, let $\mu_{t,t+k} = p_t^* - \mathrm{mc}_{t+k}^n$ denote the markup in period $t + k$ of a firm that last set its price in period t. We can rewrite (2.16) as $\mu = (1 - \beta\theta) \sum_{k=0}^{\infty} (\beta\theta)^k E_t\{\mu_{t,t+k}\}$, which yields a simple interpretation of the pricing rule: firms set prices at a level such that a (suitable) weighted average of anticipated future markups matches the optimal frictionless markup μ.

If, on one hand, firms do not adjust prices optimally each period, real marginal costs will no longer be constant. On the other hand, in a perfect foresight steady state with zero inflation, all firms will be charging their desired markup. Hence, the steady-state marginal cost, mc, will be equal to its flexible price counterpart (i.e., $-\mu$). Let $\widehat{\mathrm{mc}}_t = \mathrm{mc}_t - \mathrm{mc}$ denote the percent deviation of marginal cost from its steady-state level. We can then combine (2.15) and (2.16), and, after some algebra, obtain a simple stochastic difference equation describing the

[9] A similar assumption can be found in Rotemberg and Woodford (1999), among others.

[10] Notice that they will all be setting the same price, because they face an identical problem.

dynamics of inflation, with marginal costs as a driving force:

$$\pi_t = \beta E_t\{\pi_{t+1}\} + \lambda \widehat{\text{mc}}_t, \tag{2.17}$$

where $\lambda = \theta^{-1}(1-\theta)(1-\beta\theta)$.

Furthermore, firms' inability to adjust prices optimally every period will generally imply the existence of a wedge between output and its natural level. Let me denote that wedge by $x_t = y_t - \bar{y}_t$, and let me refer to it as the *output gap*. It follows from (2.10) that the latter will be related to marginal cost according to

$$\widehat{\text{mc}}_t = (\sigma + \varphi)x_t. \tag{2.18}$$

Combining (2.17) and (2.18) yields the familiar New Phillips Curve:

$$\pi_t = \beta E_t\{\pi_{t+1}\} + \kappa x_t, \tag{2.19}$$

where $\kappa = \lambda(\sigma + \varphi)$.

It will turn out to be convenient for the subsequent analysis to rewrite equilibrium condition (2.8) in terms of the output gap and the natural rate of interest:

$$x_t = -\frac{1}{\sigma}(r_t - E_t\{\pi_{t+1}\} - \overline{rr}_t) + E_t\{x_{t+1}\}. \tag{2.20}$$

Equations (2.19) and (2.20), together with a specification of monetary policy (i.e., of how the interest rate evolves over time), and of the exogenous processes $\{a_t\}$ and $\{g_t\}$ (which in turn determine the natural rate of interest), fully describe the equilibrium dynamics of the baseline model economy.

Having laid out the equations of the baseline sticky price model, I now turn to a discussion of some of its implications for monetary policy, inflation, and the business cycle.

3. THE NATURE OF INFLATION DYNAMICS

The nature of inflation dynamics is arguably the most distinctive feature of the NK paradigm. Yet, an important similarity with traditional Keynesian models remains on this front: as illustrated by (2.19), the evolution of inflation in the NK model is determined by some measure of the level of economic activity or, more precisely, of its deviation from some baseline level. Thus, and in contrast with classical monetary models, a change in monetary conditions (e.g., an increase in the money supply) has no direct effect on prices. Its eventual impact is only indirect, working through whatever changes in the level of economic activity it may induce. That common feature notwithstanding, there exist two fundamental differences between (2.19) and a traditional Phillips Curve. First, in the new paradigm, inflation is determined in a forward-looking manner. Second, the measure of economic activity that is the driving force behind inflation fluctuations is precisely pinned down by the theory, and may not be well approximated by conventional output-gap measures. Next I discuss those

two features in more detail, together with their empirical implications and the related evidence.

3.1. The Forward-Looking Nature of Inflation

The *traditional* Phillips Curve relates inflation to some cyclical indicator as well as its own lagged values. A simple and common specification takes the form

$$\pi_t = \pi_{t-1} + \delta \widehat{y}_t + \varepsilon_t, \tag{3.1}$$

where $\widehat{y}_t$ is the log deviation of gross domestic product (GDP) from some baseline trend (or from some measure of potential GDP) and ε_t is a random disturbance. Sometimes additional lags of inflation or detrended GDP are added. Alternative cyclical indicators may also be used (e.g., the unemployment rate). Let me emphasize two properties, which will generally hold independent of the details. First, past inflation matters for the determination of current inflation. Second, current inflation is positively correlated with past output; in other words, output leads inflation.

The previous properties stand in contrast with those characterizing the New Phillips Curve (NPC). Taking Equation (2.19) as a starting point and iterating forward yields

$$\pi_t = \kappa \sum_{k=0}^{\infty} \beta^k E_t\{x_{t+k}\}. \tag{3.2}$$

It is clear from this expression that past inflation is not, in itself, a relevant factor in determining current inflation. Furthermore, inflation is positively correlated with future output, given $\{\bar{y}_{t+k}\}$. In other words, inflation leads output, not the other way around. Thus, we see that, under the new paradigm, inflation is a forward-looking phenomenon. The intuition behind that property is clear: in a world with staggered price setting, inflation – positive or negative – arises as a consequence of price decisions by firms currently setting their prices; as made clear by (2.16), those decisions are influenced by current and anticipated marginal costs, which are in principle unrelated to past inflation.

Interestingly, many critical assessments of the NK paradigm have focused on the forward-looking nature of the inflation dynamics embedded in it. Thus, a number of authors have argued that, whereas the NPC may be theoretically more appealing, it cannot account for many features of the data that motivated the traditional Phillips Curve specification. In particular, they point out that the pattern of dynamic cross correlation between inflation and detrended output observed in the data suggests that output leads inflation, not the other way around.[11] In other words, the data appear to be more consistent with a traditional, backward-looking Phillips Curve than with the NPC. That evidence seems

[11] This point has been stressed by Fuhrer and Moore (1995), among others.

reinforced by many of the estimates of *hybrid* Phillips Curves of the form

$$\pi_t = \gamma_b \pi_{t-1} + \gamma_f E_t\{\pi_{t+1}\} + \delta(y_t - \bar{y}_t) \tag{3.3}$$

found in the literature, and that generally point to a significant (if not completely dominant) influence of lagged inflation as a determinant of current inflation.[12]

That critical assessment of the NPC has been revisited recently by Sbordone (1998), Galí and Gertler (1999; henceforth, GG), and Galí, Gertler, and López-Salido (2001; henceforth, GGL). Those authors argue that some of the existing evidence against the relevance of the NPC may be distorted by the use of detrended GDP (or similar) as a proxy for the output gap. As discussed in the next subsection, that proxy is likely to be very poor and, thus, a source of potentially misleading results. Furthermore, even if detrended GDP was highly correlated with the true output gap, the conditions under which the latter is proportional to the (current) marginal cost may not be satisfied; that would render (2.19) invalid and lead to a misspecification of the NPC formulation used in empirical work.

As a way to overcome both problems, the above-mentioned researchers have gone back one step and estimated (2.17) instead, thus taking *real marginal cost* as the (immediate) driving force underlying changes in inflation. That formulation of the inflation equation relies on weaker assumptions, because condition (2.18) is no longer required to hold. In contrast, it embeds two essential ingredients of inflation dynamics under the new paradigm: (a) the forward-looking nature of price-setting decisions and (b) their lack of synchronization (staggering).

What is most important is that they note that a theory-consistent, observable measure of average real marginal costs can be derived under certain assumptions on technology, and independent of price-setting considerations.[13] For the sake of concreteness, suppose that (a) labor productivity is exogenous, (b) firms take wages as given, and (c) there are no costs of labor adjustment. Then it is easy to show that real marginal costs will be proportional to the labor income share; it follows that $\widehat{\text{mc}}_t = \widehat{s}_t$, where $\widehat{s}_t$ denotes the percent deviation of the labor income share from its (constant) mean.

Using U.S. data on inflation and the labor income share, GG have estimated (2.17), as well as structural parameters β and θ, using an instrumental variables estimator. That evidence has recently been extended by GGL to Euroarea data. An exercise in a similar spirit has been carried out by Sbordone, using an alternative approach to estimation based on a simple goodness-of-fit criterion that seeks to minimize the model's forecast error variance, given a path for expected marginal costs.

The findings that emerge in that recent empirical work are quite encouraging for the NPC: when the latter is estimated in a way consistent with the underlying

[12] See, Chadha, Masson, and Meredith (1992) and Fuhrer (1997), among others.

[13] See Rotemberg and Woodford (1999) for a detailed discussion of alternative measures of marginals costs.

theory, it appears to fit the data much better than had been concluded by the earlier literature. Thus, all the parameter estimates have the predicted sign and show plausible values. In particular, estimates of parameter θ imply an average price duration of about one year, which appears to be roughly consistent with the survey evidence.[14]

The GG and GGL papers also provide an extension of the baseline theory underlying the NPC to allow for a constant fraction of firms that set prices according to a simple, backward-looking rule of thumb. The remaining firms set prices in a forward-looking way, as in the baseline sticky price model. The reduced-form inflation equation that results from the aggregation of pricing decisions by both types of firms takes a hybrid form similar to (3.3), with a measure of marginal cost replacing the output gap, and with coefficients γ_b and γ_f being a function of all structural parameters (now including the fraction of firms that are backward looking). The findings there are also quite encouraging for the baseline NPC: although backward-looking behavior is often statistically significant, it appears to have limited quantitative importance. In other words, although the baseline pure forward-looking model is rejected on statistical grounds, it is still likely to be a reasonable first approximation to the inflation dynamics of both Europe and the United States.[15]

3.2. The Nature of the Output Gap

According to the NPC paradigm, inflation fluctuations are associated with variations in the output gap, that is, in the deviation of output from its level under flexible prices.[16] The output gap and its volatility also play an important role in welfare evaluations: as shown in Rotemberg and Woodford (1997), the variance of the output gap is one of the key terms of a second-order approximation to the equilibrium utility of the representative consumer, in the context of a model similar to the one sketched herein.

The concept of output gap associated with the NK model is very different from the one implicit in most empirical applications. In the latter, the concept of output gap used would be better characterized as a measure of detrended output, that is, deviations of log GDP from a *smooth* trend. That trend is computed by using one of a number of available procedures, but the main properties of the resulting series do not seem to hinge critically on the exact procedure used. This is illustrated in Figure 5.1, which plots three output-gap series for the U.S. economy commonly used in empirical work. Those gap measures correspond to three alternative estimates of the trend: (a) a fitted quadratic function of time, (b) a Hodrick–Prescott filtered series, and (c) the Congressional Budget

[14] See Taylor (1999) for an overview of that evidence.

[15] Interestingly, as shown in GGL, the backward-looking component appears to be even less important in Europe than in the United States.

[16] As should be clear from the derivation of (2.19), that relationship is not a primitive one. It arises from the proportionality between the output gap and markups (or real marginal costs) that holds under some standard, though by no means general, assumptions.

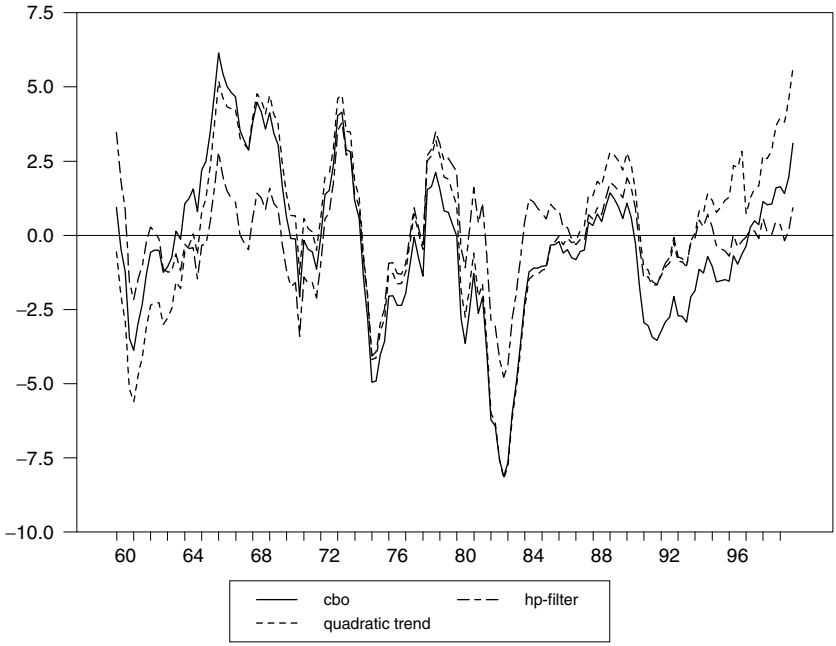


Figure 5.1. Three conventional output-gap measures.

Office's estimate of potential output. The fact that the implied trend is a very smooth series has two implications. First, the bulk of the fluctuations in output at business cycle frequencies are attributed to fluctuations in the output gap. Second, the correlation among the three output-gap measures is very high.

But, as argued in GG and Sbordone, the use of detrended GDP as a proxy for the output gap does not seem to have any theoretical justification. In effect, that approach implicitly assumes that the natural level of output $\{\bar{y}_t\}$ can be represented as a smooth function of time. Yet, the underlying theory implies that any shock (other than monetary shocks) may be a source of fluctuations in that natural level of output; as a result, the latter may be quite volatile.[17] In fact, one of the tenets of the RBC school was that the bulk of the business cycle in industrial countries could be interpreted as the equilibrium response of a *frictionless* economy to technology and other real shocks; in other words, to fluctuations in the *natural* level of output!

Under the assumptions made in Section 2, and as shown in Equation (2.18), the true output gap is proportional to deviations of real marginal cost from the steady state. Hence, a measure of real marginal cost can be used to approximate (up to a scalar factor) the true, or model-based, output gap. Figure 5.2 displays a time series for the U.S. output gap, defined as $x_t = 0.5\widehat{s}_t$. Notice that this is consistent, for example, with parameter settings $\sigma = 1$ and $\varphi = 1$; those values

[17] See, e.g., Rotemberg and Woodford (1999) for an illustration of this point in the context of a calibrated version of a sticky price model.

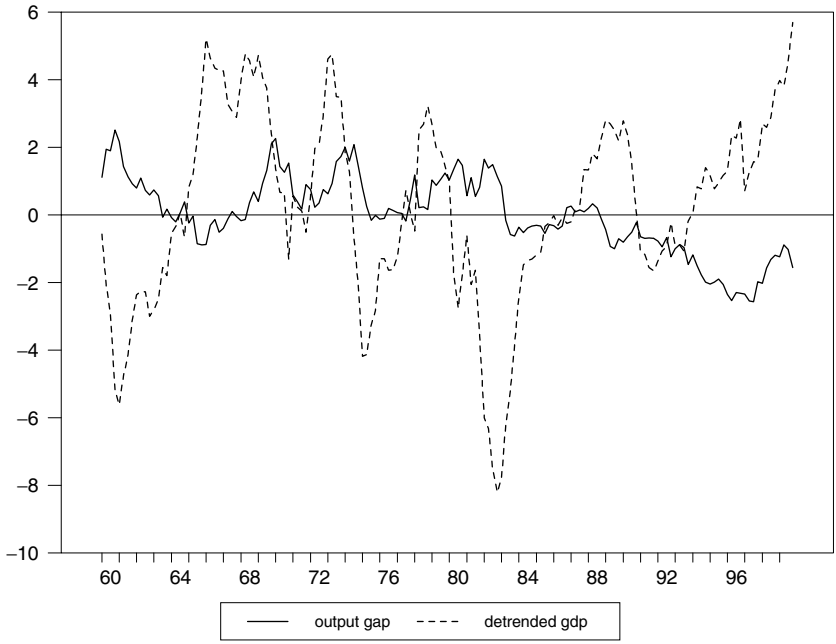


Figure 5.2. Model-based output gap vs. detrended GDP.

arguably fall within a reasonable range. In addition, Figure 5.2 also shows the deviation of log GDP from a fitted quadratic trend, a popular proxy for the output gap in empirical applications. Let me not emphasize here the apparent differences in volatility between the two series, because the model pins down the output gap only up to a scale factor (determined by the choice of settings for σ and φ). Instead I want to focus on their comovement: if detrended GDP was a good proxy for the output gap, we should observe a strong positive comovement between the two series. But a look at Figure 5.3 makes it clear that no obvious relationship exists; in fact, the contemporaneous correlation between them turns out to be slightly negative. As argued in Galí and Gertler (1999), the previous finding calls into question the validity of empirical tests of the NPC that rely on detrended GDP as a proxy for the output gap, including informal assessments based on the patterns of cross correlations between that variable and inflation.

4. THE EFFECTS AND TRANSMISSION OF SHOCKS

Having laid out the baseline NK model and discussed some of its most distinctive elements, I turn to the examination of some of its predictions regarding the effects of some aggregate shocks on the economy.

Much of the quantitative analysis that follows relies on a baseline calibration of the model, though a number of variations from it are also considered. In the

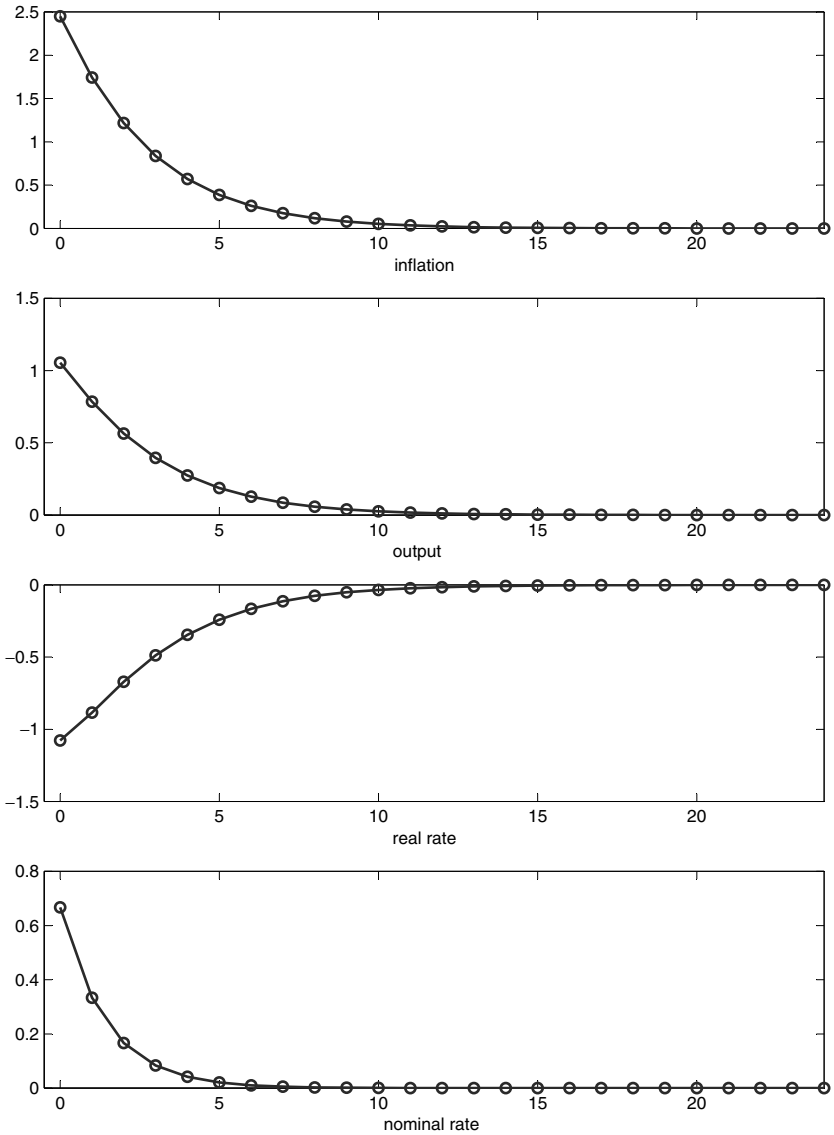


Figure 5.3. Dynamic responses to a monetary shock.

baseline calibration, I assume a log utility for consumption, which corresponds to $\sigma = 1$. This is a standard assumption, and one that would render the model consistent with a balanced growth path if secular technical progress were introduced. I also set $\varphi = 1$, which implies a unit wage elasticity of labor supply. The baseline value for the semi-elasticity of money demand with respect to the (quarterly) interest rate, η, is set to unity. This is roughly consistent with an

interest elasticity of 0.05 found in empirical estimates and used in related work.[18] The baseline choice for θ is 0.75. Under the Calvo formalism, that value implies an average price duration of one year. This appears to be in line with econometric estimates of θ, as well as survey evidence.[19] The elasticity of substitution ε is set to 11, a value consistent with a 10 percent markup in the steady state. Finally, I set $\beta = 0.99$, which implies an average annual real return of about 4 percent.

4.1. Monetary Policy Shocks

As is well known, the presence of nominal rigidities is a potential source of nontrivial real effects of monetary policy shocks. This is also the case for the baseline NK model, where firms do not always adjust the price of their good when they receive new information about costs or demand conditions.

What are the real effects of monetary policy shocks in this framework? How are they transmitted? To focus attention on these issues, we abstract momentarily from nonmonetary shocks by assuming $a_t = g_t = 0$, for all t. Without loss of generality I also set $\bar{y}_t = 0$, all t.

Solving (2.8) forward, we obtain

$$y_t = -\frac{1}{\sigma} \sum_{k=0}^{\infty} E_t\{r_{t+k} - \pi_{t+k+1} - \rho\}. \tag{4.1}$$

We see that, in the NK model, exogenous interventions by the monetary authority will have an effect on output only to the extent that they influence current or future expected short-term real interest rates or, equivalently, under the expectations hypothesis of the term structure, only if they affect the current long-term real rate.

To understand how the transmission works, I specify monetary policy by assuming an exogenous path for the growth rate of the money supply, given by the stationary process

$$\Delta m_t = \rho_m \Delta m_{t-1} + \varepsilon_t^m, \tag{4.2}$$

where $\rho_m \in [0, 1)$. Under that assumption, the equilibrium dynamics of the baseline NK model are described by the stationary system

$$\begin{bmatrix} 1 + \frac{1}{\sigma\eta} & 0 & 0 \\ -\kappa & 1 & 0 \\ 0 & -1 & 1 \end{bmatrix} \begin{bmatrix} y_t \\ \pi_t \\ m_{t-1} - p_{t-1} \end{bmatrix} = \begin{bmatrix} 1 & \frac{1}{\sigma} & \frac{1}{\sigma\eta} \\ 0 & \beta & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} E_t\{y_{t+1}\} \\ E_t\{\pi_{t+1}\} \\ m_t - p_t \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ -1 \end{bmatrix} \Delta m_t. \tag{4.3}$$

[18] See, e.g., Chari, Kehoe, and McGrattan (1996).

[19] Taylor (1999) summarizes the existing survey evidence.

As a baseline setting for ρ_m, I choose 0.5, a value consistent with the estimated autoregressive process for M1 in the United States. The estimated standard deviation of the money shock, denoted by σ_m, is approximately 0.01. For convenience, I set $\sigma_m = 1$, which requires that the units of all variables be interpreted as percentage points or percent deviations.[20]

Figure 5.3 displays the dynamic responses of output, inflation, and both nominal and (ex ante) real rates to a one-standard-deviation money supply shock, under the baseline calibration. For ease of interpretation, the rates of inflation and interest rate displayed in the figure have been annualized (the calibration is based on quarterly rates, however).

I would like to highlight two features of the responses shown in Figure 5.3. First, they suggest that, in the simple sticky price model considered here, a "typical" monetary shock has strong, and highly persistent, effects on output. Second, under the baseline calibration, a monetary expansion is predicted to raise the nominal rate; in other words, the calibrated model does *not* predict the existence of a liquidity effect. Next I briefly discuss each of these properties in turn.

4.1.1. The Effects of Money on Output: Strength and Persistence

What level of GNP volatility can be accounted for by the basic NK model, when shocks to an exogenous money supply process (calibrated in accordance with postwar U.S. data) are the only source of fluctuations? For our baseline calibration, the answer to the previous question is a surprisingly large value: 2.1 percent. That value is significantly above the estimated standard deviation of detrended U.S. GDP in the postwar period.[21] Interestingly, and despite their focus on the persistence of the shocks, Chari, Kehoe, and McGrattan (2000) document a dual result in the context of the Taylor-type model: They calibrate σ_m in order to match the volatility of output, leading them to set at a value well below the estimated one.[22]

Although the previous exercise is useful for pointing out the powerful real effects of changes in the money supply, there are many reasons not to take it too literally. For one, the estimated variance of money supply shocks is likely to overstate the true volatility of the unexpected component of money, because, by construction, specification (4.2) ignores the existence of any endogenous component of variations in the money supply.[23] That notwithstanding, Figure 5.3 makes clear that the effects of monetary policy on output are far from negligible: on impact, a 1 percent increase in the money supply raises output by more than 1 percent, whereas the implied increase in the price level is about 2.4 percent (annualized).

[20] Cooley and Hansen (1989), Walsh (1998), and Yun (1996) justify the choice of that calibration.

[21] Stock and Watson (1999) report a standard deviation of 1.66 percent for the period 1953–96, using a bandpass filter to isolate cyclical fluctuations. Other estimates in the literature are similar.

[22] See also Walsh (1998) and Yun (1996) for a similar result.

[23] The analogy with the calibration of technology changes based on an estimated process for the Solow residual seems appropriate.

In addition to the large output effects of money discussed herein, the baseline model also implies that such effects are quite persistent. That property is apparent in the impulse response of output displayed in Figure 5.3. In particular, the half-life of that output response under the baseline calibration is 3.2 quarters.[24]

4.1.2. The Presence (or Lack Thereof) of a Liquidity Effect

As shown in Figure 5.3, under the baseline calibration of the NK model, a monetary expansion raises the nominal rate. Hence, and in contrast with a textbook model, the calibrated model does not predict the existence of a liquidity effect. Still, that feature does not prevent monetary policy from transmitting its effects through an interest rate channel: as shown in the same figure, the (ex ante) real rate declines substantially when the monetary expansion is initiated, remaining below its steady-state level for a protracted period. As (4.1) makes clear, it is that persistent decline that induces the observed expansion in aggregate demand and output.

The absence of a liquidity effect is not, however, a robust feature of the NK model. Yet, and as discussed in Christiano, Eichenbaum, and Evans (1997) and in Andrés, López-Salido, and Vallés (1999), standard specifications of preferences and the money growth process tend to rule out a liquidity effect. To understand the factors involved, notice that the interest rate can be written as[25]

$$r_t = \left(\frac{\sigma - 1}{1 + \eta}\right) \sum_{k=1}^{\infty} \left(\frac{\eta}{1 + \eta}\right)^{k-1} E_t\{\Delta y_{t+k}\} + \left[\frac{\rho_m}{1 + \eta(1 - \rho_m)}\right] \Delta m_t. \tag{4.4}$$

Under the baseline calibration, we have $\sigma = 1$; in that case, (4.4) implies that the nominal rate will be proportional to the expected growth rate of money, $\rho_m \Delta m_t$. To the extent that money growth is positively serially correlated (as is the case in our baseline calibration), the nominal rate will necessarily increase in response to a monetary expansion.

Under what conditions can the liquidity effect be restored? Notice that, to the extent that a monetary expansion raises output on impact, the term

$$\sum_{k=1}^{\infty} \left(\frac{\eta}{1 + \eta}\right)^{k-1} E_t\{\Delta y_{t+k}\}$$

[24] The previous result contrasts with the findings of Chari et al. (1996), who stress the difficulty in generating significant effects of money on output beyond the duration of price (which is deterministic in their framework).

[25] This is in order to derive that expression first difference (2.5) and combine the resulting expression with (2.20); some algebraic manipulation then yields the expression in the text.

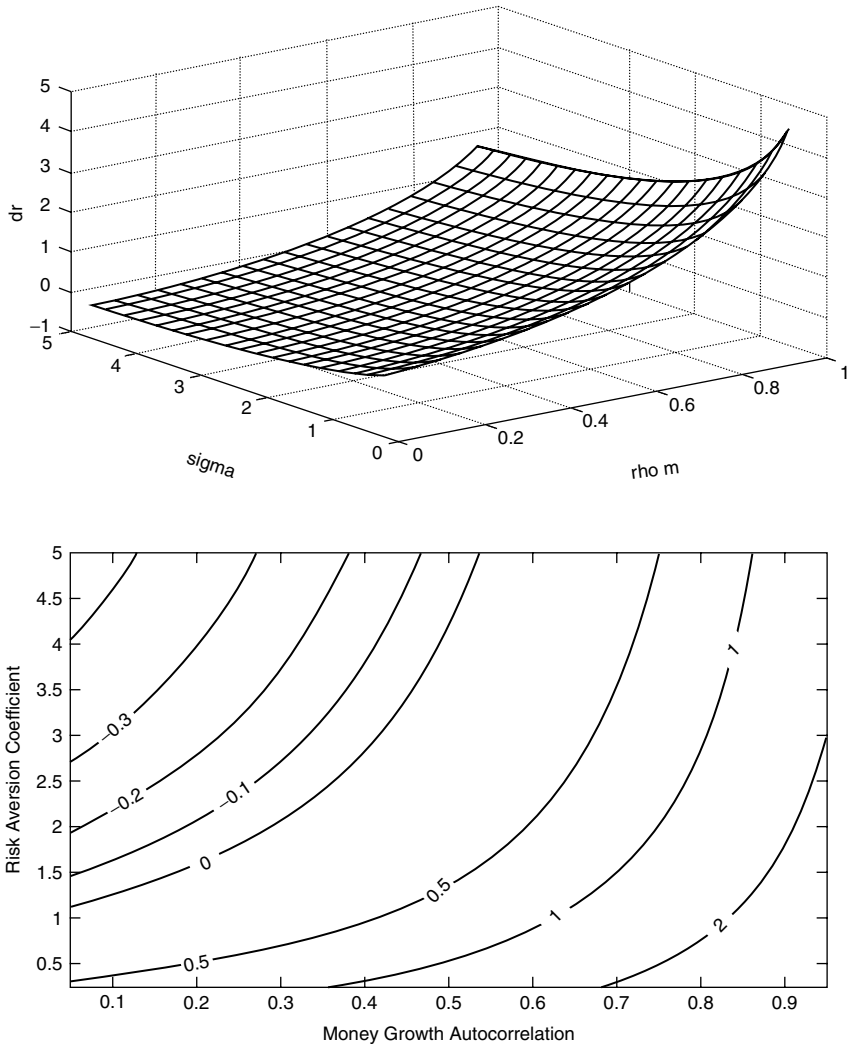


Figure 5.4. Monetary shocks and the liquidity effect.

will generally be negative.[26] Accordingly, the presence of a liquidity effect requires a sufficiently high risk-aversion parameter σ (for any ρ_m), or, given $\sigma > 1$, a sufficiently low money growth autocorrelation ρ_m. The previous trade-off is clearly illustrated in Figure 5.4, which displays the impact effect on the nominal rate of a unit monetary shock, as a function of σ and ρ_m. The bottom

[26] Long-run neutrality of money implies that $\lim_{k\to\infty} E_t\{y_{t+k}\} = 0$. Hence, the sign of the summatory in (4.4) will be positive if (a) the output's reversion to its initial level is monotonic (as in the impulse response displayed in Figure 5.5), and/or (b) if η is sufficiently large.

graph shows a sample of loci of σ and ρ_m configurations associated with a given interest rate change. Hence, the zero locus represents the "liquidity effect frontier": Any $\{\sigma, \rho_m\}$ combination above and to the left of that locus will be associated with the presence of a liquidity effect. We see that, under the baseline calibration ($\rho_m = 0.5$), a risk-aversion parameter of size slightly above 4 is necessary to generate a liquidity effect.

4.2. Technology Shocks

Proponents of the RBC paradigm have claimed a central role for exogenous variations in technology as a source of observed economic fluctuations. In contrast, the analysis of models with nominal rigidities has tended to emphasize the role of demand and, in particular, monetary disturbances as the main driving forces behind the business cycle.

Recently, however, a number of papers have brought attention to a surprising aspect of the interaction between sticky prices and technological change. In particular, Galí (1999) and Basu, Fernald, and Kimball (1998; henceforth, BFK) have made the following observation: in a model with imperfect competition and sticky prices, a favorable technology shock is likely to induce a short-run decline in employment, as long as the response of the monetary authority falls short of full accommodation.[27] That prediction is illustrated in Figure 5.5, where the dynamic responses of inflation, the output gap, output, and employment to a 1 percent permanent increase in productivity are displayed, using the baseline model developed herein under the assumption of a constant money supply. Figure 5.6 examines the robustness of that prediction to changes in the degree of price rigidities and the risk-aversion parameter, by displaying the response of employment on impact for a range of values of those parameters, as well as the corresponding contour plots. Hence we see that although a negative response of employment to a favorable technology shock is not a necessary implication of the model, that prediction appears to hold for a very large subset of the parameter values considered. More specifically, employment is seen to increase in response to a positive technology shock only when a low risk-aversion parameter coexists with little nominal rigidities (i.e., the southwest corner of the figure).

The intuition behind that result can easily be grasped by considering the case of an interest-inelastic money demand, so that $y_t = m_t - p_t$ holds in equilibrium. Assume, for the sake of argument, that the money supply remains unchanged in the wake of a technology shock. Notice also that even though all firms will experience a decline in their marginal cost, only a fraction of them will adjust their prices downward in the short run. Accordingly, the aggregate price level will decline, and aggregate demand will rise, less than proportionally

[27] Not surprisingly, the pattern of response of employment and any other variable will depend on the systematic response of the monetary authority to those shocks, as argued in Dotsey (1999).

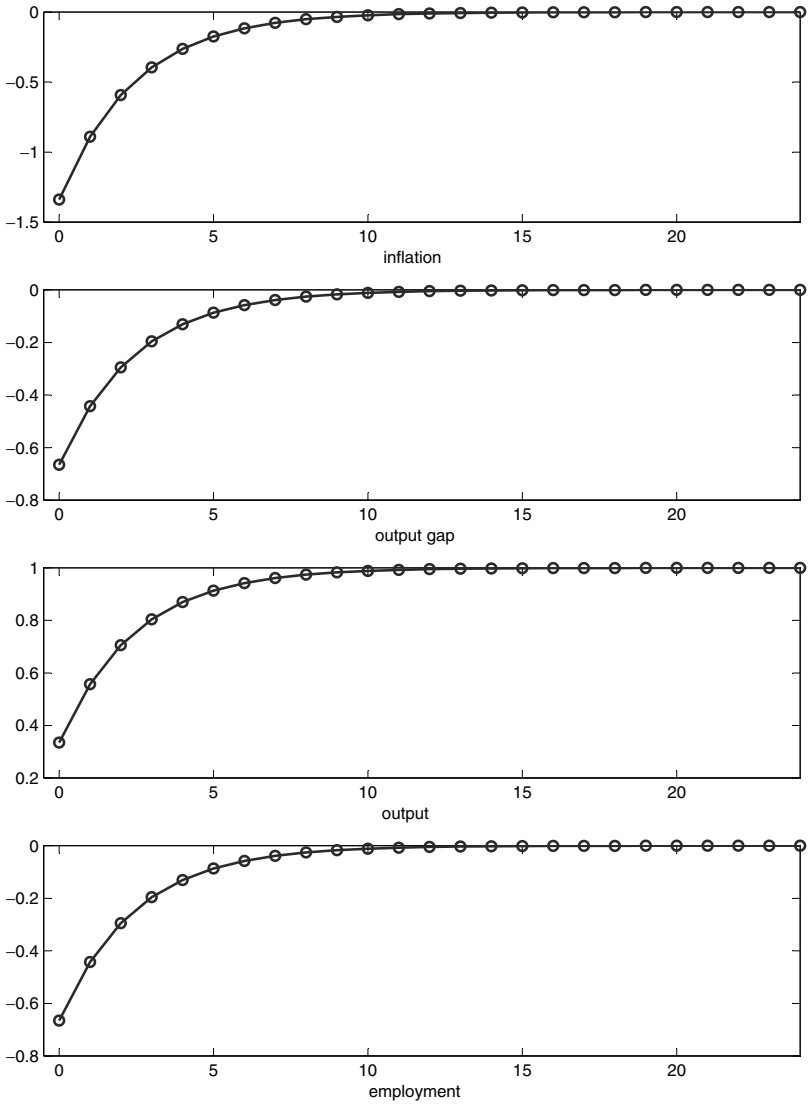


Figure 5.5. Dynamic responses to a technology shock.

to the increase in productivity. That, in turn, induces a decline in aggregate employment.[28]

[28] BFK's model allows for the possibility of a short-run decline in output after a positive technology shock. That outcome can be ruled out in the baseline model considered here, under the assumption of a constant money supply and $\sigma = 1$. In that case the nominal rate remains unchanged, see Equation (4.4), and output has to move in the opposite direction from prices (which go down).

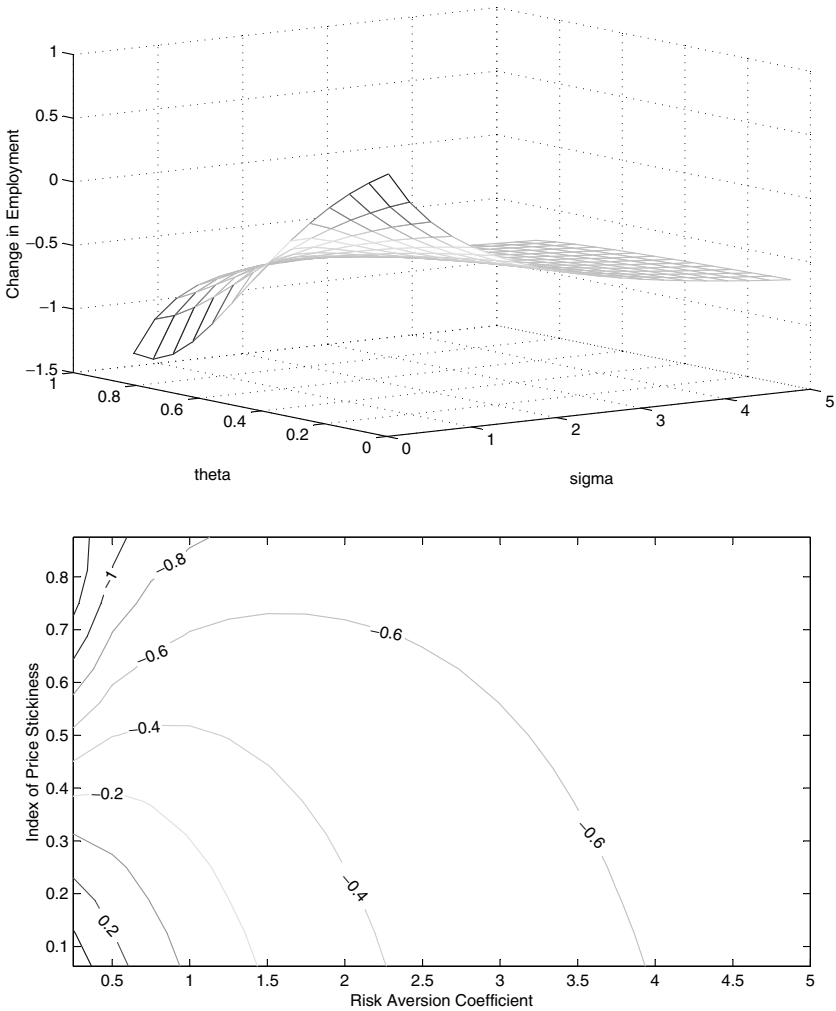


Figure 5.6. Technology shocks and employment.

The previous characterization of the economy's response to a positive technology shock is clearly at odds with some central implications of the standard RBC model. The latter's prediction of a positive short-run comovement among productivity, output, and employment in response to technology shocks lies at the root of the ability of an RBC model to replicate some central features of observed aggregate fluctuations, while relying on exogenous variations in technology as the only (or, at least, the dominant) driving force.

But, how does the actual economy respond to technology shocks? Which of the two competing frameworks does it favor? Galí (1999) and BFK (1998) provide some evidence pertaining to this matter, by estimating the responses of

a number of variables to an identified technology shock. Although the approach to identification is very different in the two cases, the results that emerge are similar: In response to a positive technology shock, labor productivity rises more than output, while employment shows a persistent decline. Hence, conditional on technology as a driving force, the data point to a negative correlation between employment and productivity, as well as between employment and output. Both observations call into question the empirical relevance of the mechanism through which aggregate variations in technology are transmitted to the economy in the basic RBC model. What is perhaps most important, and independent of the reference model, is that they raise serious doubts about the quantitative significance of technology shocks as a source of aggregate fluctuations in industrialized economies.

5. THE DESIGN OF MONETARY POLICY

The previous section looked at the effects of exogenous changes in the money supply in the context of a calibrated sticky price model. The usefulness of that sort of analysis is twofold. First, it helps us understand the way changes in monetary policy are transmitted to a number of macroeconomic variables in the model of reference. Second, it allows for an empirical evaluation of the underlying model, through a comparison of the estimated responses to an exogenous monetary shock against the model's predictions regarding the effects of a monetary intervention that corresponds to the experiment observed in the data.[29] Yet, the limitations of such a specification of monetary policy are by now well understood. For one, the common practice of modern central banks does not involve the use of the quantity of money as a policy instrument, and only very seldom as an intermediate target. Furthermore, the assumption of an exogenous random process for the money supply (or any other policy instrument, for that matter), although convenient for certain purposes, can hardly be viewed as a plausible one, because it is equivalent to modeling monetary policy as a process of randomization over the possible values of a policy instrument. This is clearly at odds with even a casual observation of how central banks conduct monetary policy.

Instead, much recent research in monetary economics, both theoretical and empirical, has deemphasized the analysis of monetary shocks and its effects, and turned instead its focus on the endogenous component of monetary policy.[30] There are several natural questions that one can ask in that context. What are

[29] See Christiano et al. (1998) for a discussion of that methodological approach.

[30] That shift in emphasis may not be unrelated to a common finding in the structural VAR literature: identified exogenous monetary policy disturbances account for only a relatively small portion of the observed fluctuations in output and other macroeconomic variables, including monetary instruments and aggregates. An important component of the latter's variations must hence be attributed to the systematic reaction of the monetary authority to macroeconomic developments, that is, to the endogenous component of monetary policy. See Christiano et al. (1999) for a recent overview of that literature.

the goals of monetary policy? How should monetary policy be conducted in order to achieve those goals? In particular, how should the monetary authority respond to different shocks? What are the losses from following simple rules, which depart from the optimal one?

It is beyond the scope of this paper to attempt to provide a general treatment of such important questions. Instead, I use the baseline sticky price model developed herein to illustrate how some of those questions can be addressed with the tools of modern monetary theory. As a way to keep things manageable, and to focus on the role of nominal rigidities in the design of monetary policy, most of the analysis maintains the assumption that the presence of such rigidities is the only distortion left for the monetary authority to correct.

5.1. The Goals of Monetary Policy

In setting up the problem facing the monetary authority, we adopt the natural assumption that it acts as a benevolent policymaker; that is, it seeks to maximize the utility of the representative household. That optimization problem is subject to some economywide resource constraints, as well as institutional and/or informational constraints that may limit the sort of interventions that the monetary authority can undertake.

In practice, the optimizing policy maker will seek to eliminate (or, at least, to offset as much as possible) any distortions that may exist in the economy. Models of monetary economies with staggered price setting found in the literature would typically have as many as four distortions, briefly described next.[31]

A first distortion results from agents' need or desire to allocate part of their wealth to (non-interest-bearing) monetary assets, given the transaction services provided by the latter. The existence of a private opportunity cost of holding such monetary balances, coexisting with a zero social cost of producing them, generates a *transactions–monetary distortion*. The elimination of that distortion requires that the nominal interest rate is set to zero (the Friedman rule), in order to equate private and social costs.[32] The average level of inflation associated with a zero nominal rate is given by minus the steady-state equilibrium real rate ($\pi = -\rho$, in the framework herein). Hence, policies that seek to implement the Friedman rule will generally involve a steady decline in the price level.

A second distortion results from the existence of imperfect competition in the goods market. That feature is reflected in prices that are, on average, above marginal cost. In equilibrium, that property makes the marginal rate of substitution between consumption and labor (the real wage) differ from their corresponding marginal rate of transformation (the marginal product of labor).

[31] See Khan, King, and Wolman (2000) and Woodford (1999c) for a detailed analysis of the role played by each distortion in the design of monetary policy.

[32] See Correia and Teles (1999) for a modern treatment of the optimality of the Friedman rule in classical monetary models, under a variety of assumptions.

The existence of such a *static or average markup distortion* is often invoked to rationalize an objective function for the policy maker that penalizes deviations of the output gap x_t from a positive target $x^* > 0$. The latter corresponds to the efficient level of activity (i.e., the one that would be obtained under perfect competition). Thus, and in the absence of an employment subsidy ($v = 0$), the output-gap target in the model of Section 2 would be given by $x^* = \mu/(\sigma + \varphi)$. The steady-state rate of inflation that would be associated with that output gap can easily be derived by using (2.19), and is approximately given by $\pi \simeq (\lambda\mu/\rho) > 0$. Thus, using monetary policy to offset the distortion created by the existence of market power would require a departure from the Friedman rule, and would thus conflict with attempts to reduce the monetary distortion.[33]

Notice that the two distortions just discussed would be operative even in the limiting case of full price flexibility; in other words, they are *not* related to the presence of nominal rigidities. In the subsequent analysis I choose to ignore the previous distortions, and instead I focus on those resulting from the presence of sticky prices and staggered price setting. This is equivalent to taking the welfare gains from holding higher real balances to be small[34] (relative to the utility derived from consumption of goods and leisure), and to assume that the static market power distortion is exactly offset by means of an employment subsidy of the right size (financed through lump sum taxes). In such a world there would remain at least two additional distortions, both directly related to the presence of sticky prices and the staggered nature of price setting. On one hand, firms' inability to adjust prices each period in the face of shocks will lead to persistent deviations of markups from their frictionless level (equivalently, it will generate inefficient fluctuations in the output gap). Let me refer to that source of inefficiency as the *dynamic markup distortion*. On the other hand, the lack of synchronization in price adjustments will generally imply the coexistence of different prices (and, hence, different quantities produced and consumed) for goods that enter consumers' preferences symmetrically and that have a one-to-one marginal rate of transformation. That *relative price distortion* induces an inefficiency in the allocation of resources that remains even in the absence of markup fluctuations.[35]

Interestingly, it turns out that, in the model considered here, the two distortions associated with the presence of sticky prices (a) do *not* generate a policy trade-off, and (b) can be fully corrected through an appropriate monetary policy. The optimal policy requires that the markup of all firms is fully stabilized

[33] See King and Wolman (1996) for a comprehensive analysis of that long-run trade-off, as well as a full characterization of the steady state of the Calvo model. Notice that, under our baseline calibration, and assuming a frictionless net markup of 10 percent ($\mu = 0.1$), the rate of inflation consistent with the efficient level of activity would be about 80 percent per quarter.

[34] In models in which money is an argument in the utility function, that assumption is equivalent to letting the coefficient on real balances approach zero.

[35] Thus, in a steady state with nonzero inflation (and constant average markups), only a fraction of individual prices will be adjusted each period, generating dispersion in relative prices.

at its flexible price level. That stabilization will be attained only if the price level is fully stabilized (permanent zero inflation), as implied by (3.2). In that environment, firms that have an opportunity to readjust their prices choose not to do so; that, in turn, requires that they all charge a common optimal markup μ and, hence, share identical prices and quantities (implying the absence of a relative price distortion). In other words, the constraint on firms' ability to adjust prices becomes nonbinding, and the flexible price allocation is restored. Along with it, the fact that no firm has an incentive to change its price implies that the path of the aggregate price level is perfectly flat (zero inflation).

In that context, and as shown in Woodford (1999c), the period utility losses resulting from deviations from the flexible price allocation can be approximated by means of the period loss function[36]:

$$L_t = \frac{U_c C}{2}\left((\sigma+\varphi)\, E_t\{x_t^2\} + \frac{\varepsilon}{\lambda}\, E_t\{\pi_t^2\}\right). \tag{5.1}$$

Hence, the expected loss of utility resulting from departures from the optimal allocation, expressed as a fraction of steady-state consumption, is approximately given as

$$\frac{1}{2}\left[(\sigma+\varphi)\mathrm{var}(x_t) + \frac{\varepsilon}{\lambda}\mathrm{var}(\pi_t)\right]. \tag{5.2}$$

Next I derive and characterize the monetary policy strategy that would minimize those losses, and I discuss some issues related to its implementation.

5.2. Optimal Monetary Policy

A monetary authority seeking to minimize the loss function (5.2) does not face any trade-off: It will find it possible to fully stabilize both inflation and the output gap. Thus, the optimal policy requires that

$$x_t = \pi_t = 0,$$

all t. The resulting allocation under that policy replicates the (efficient) flexible price equilibrium allocation. Given the price stability requirement, the path for the nominal rate consistent with the optimal policy will correspond to that of the real rate in the flexible price equilibrium. Hence, and given (2.14), the optimal policy implies

$$\begin{aligned} r_t &= \overline{rr}_t \\ &= \rho + \sigma\psi_a\rho_a\Delta a_t + \sigma(1-\psi_g)(1-\rho_g)g_t, \end{aligned} \tag{5.3}$$

where we recall that $\psi_a = (1+\varphi)/(\sigma+\varphi)$, and $\psi_g = \sigma/(\sigma+\varphi)$. The intuition underlying the optimality of that interest rate response can easily be grasped

[36] See Rotemberg and Woodford (1999) and Woodford (1999c) for a derivation of the approximated welfare loss function under alternative and more general assumptions, respectively.

by considering the adjustment of consumption to both technology and fiscal shocks in the flexible price case; see (2.12). Thus, an expansionary fiscal shock lowers consumption on impact, with the reversion to its initial level requiring a higher interest rate. In contrast, as long as there is positive serial correlation in productivity growth ($\rho_a > 0$), a positive technology shock leads to a gradual adjustment of consumption to its new, higher plateau; supporting that response pattern also requires a higher interest rate.[37]

5.2.1. Implementation

Interestingly, however, (5.3) cannot be interpreted as a monetary policy rule that the central bank could follow mechanically, and that would guarantee that the optimal allocation is attained. To see this, notice that, after plugging (5.3) into (2.20), we can represent the equilibrium dynamics by means of the difference equation

$$\begin{bmatrix} x_t \\ \pi_t \end{bmatrix} = \mathbf{A_O} \begin{bmatrix} E_t\{x_{t+1}\} \\ E_t\{\pi_{t+1}\} \end{bmatrix}, \tag{5.4}$$

where

$$\mathbf{A_O} = \begin{bmatrix} 1 & \sigma^{-1} \\ \kappa & \beta + \kappa\sigma^{-1} \end{bmatrix}.$$

Clearly, $x_t = \pi_t = 0$, for all t, constitutes a solution to (5.4). Yet, a necessary and sufficient condition for the uniqueness of such a solution in a system with no predetermined variables such as (5.4) is that the two eigenvalues of $\mathbf{A_O}$ lie inside the unit circle.[38] It is easy to check, however, that such a condition is not satisfied in our case. More precisely, although both eigenvalues of $\mathbf{A_O}$ can be shown to be real and positive, only the smallest one lies in the [0, 1] interval. As a result, there exists a continuum of solutions in a neighborhood of (0, 0) that satisfy the equilibrium conditions (local indeterminacy). Furthermore, one cannot rule out the possibility of equilibria displaying fluctuations driven by self-fulfilling revisions in expectations (stationary sunspot fluctuations).

That indeterminacy problem can be avoided, and the uniqueness of the equilibrium allocation restored, by having the central bank follow a rule that would make the interest rate respond to inflation and/or the output gap were those variables to deviate from their (zero) target values. More precisely, suppose that the central bank commits itself to following the rule

$$r_t = \overline{rr}_t + \phi_\pi \pi_t + \phi_x x_t. \tag{5.5}$$

[37] See Galí, López-Salido, and Vallés (2000) for evidence of an efficient response by the Fed to technology shocks during the Volcker–Greenspan period, which contrasts with the inefficient response observed during the pre-Volcker period.

[38] See, for example, Blanchard and Kahn (1980).

In that case, the equilibrium is described by a stochastic difference equation such as (5.4), with $\mathbf{A_O}$ replaced with

$$\mathbf{A}_T = \Omega \begin{bmatrix} \sigma & 1-\beta\phi_\pi \\ \sigma\kappa & \kappa+\beta(\sigma+\phi_x) \end{bmatrix},$$

where $\Omega = 1/(\sigma + \phi_x + \kappa\phi_\pi)$. If we restrict ourselves to nonnegative values of ϕ_π and ϕ_x, then a necessary and sufficient condition for $\mathbf{A}_T$ to have both eigenvalues inside the unit circle, thus implying uniqueness of the (0,0) solution to (5.4), is given by[39]

$$\kappa(\phi_\pi - 1) + (1-\beta)\phi_x > 0. \tag{5.6}$$

Notice that, once uniqueness is restored, the term $\phi_\pi \pi_t + \phi_x x_t$ appended to the interest rate rule vanishes, implying that $r_t = \overline{rr}_t$, all t. Hence, we see that stabilization of the output gap and inflation requires a credible threat by the central bank to vary the interest rate sufficiently in response to any deviations of inflation and/or the output gap from target; yet, the very existence of that threat makes its effective application unnecessary.

5.2.2. Discussion

A common argument against the practical relevance of a monetary policy rule such as (5.5) stresses the fact that its implementation requires having far more information than that available to actual central banks.[40] For one, the specific form of the optimal policy rule is not robust to changes in some of the model characteristics; its correct application thus hinges on knowledge of the true model and of the values taken by all its parameters. In addition, it requires that the central bank be able to observe and respond (contemporaneously) to the realizations of the different shocks (a_t and g_t, in the present model). This is not likely to be the case in practice, given the well-known problems associated with measurement of variables such as total factor productivity, not to mention the practical impossibility of detecting exogenous shifts in some parameters that may be unobservable by nature (e.g., parameters describing preferences).[41]

The practical difficulties in implementing optimal rules have led many authors to propose a variety of simple rules as possible alternatives, and to evaluate their desirability in the context of one or more models. A large number of recent papers have sought to analyze the properties and desirability of many

[39] See, for example, Bullard and Mitra (1999).

[40] See Blinder (1998) for a discussion of the practical complications facing central bankers in the design and implementation of monetary policy.

[41] Given the inherent difficulties in measuring the output gap, the previous argument would also seem to apply to the central bank's need to respond to that variable in order to avoid the indeterminacy problem. But it is clear from (5.6) that the equilibrium is unique even if $\phi_x = 0$, as long as $\phi_\pi > 1$.

such rules.[42] In the next subsection I describe three of them and examine their properties in the context of the baseline sticky price model developed herein.

5.3. Simple Policy Rules

Here I embed three alternative simple rules in the baseline sticky price model and analyze their basic properties. The following three rules are considered in turn: a Taylor rule, a constant money growth rule, and an interest rate peg.

5.3.1. A Simple Taylor Rule

Let us assume that the central bank follows the rule

$$r_t = \rho + \phi_\pi \pi_t + \phi_x x_t; \tag{5.7}$$

that is, the nominal rate responds systematically to the contemporaneous values of inflation and the output gap. This is a version of the rule put forward by John Taylor as a good characterization of U.S. monetary policy, and analyzed in numerous recent papers.[43]

Combining (5.7) with (2.19) and (2.20), one can represent the equilibrium dynamics with the following system:

$$\begin{bmatrix} x_t \\ \pi_t \end{bmatrix} = \mathbf{A}_T \begin{bmatrix} E_t\{x_{t+1}\} \\ E_t\{\pi_{t+1}\} \end{bmatrix} + \mathbf{B}_T(\overline{rr}_t - \rho),$$

where $\mathbf{A}_T$ and $\overline{rr}_t$ are defined as just given, and $\mathbf{B}_T = \Omega\,[1, \kappa]'$. As long as ϕ_π and ϕ_x satisfy condition (5.6), the previous system has a unique stationary solution, which can be written as

$$\begin{bmatrix} x_t \\ \pi_t \end{bmatrix} = \omega_a[\mathbf{I} - \rho_a \mathbf{A}_T]^{-1}\mathbf{B}_T \Delta a_t + \omega_g[\mathbf{I} - \rho_g \mathbf{A}_T]^{-1}\mathbf{B}_T g_t,$$

where $\omega_a = \sigma\psi_a\rho_a$ and $\omega_g = \sigma(1 - \psi_g)(1 - \rho_g)$.[44]

[42] See, for example, the contributions by several authors contained in the Taylor (1999) volume.

[43] See Taylor (1993, 1999) and Judd and Rudebusch (1988). Clarida, Galí, and Gertler (1998, 2000) estimate a forward-looking version of that rule, in which the interest rate is assumed to respond to anticipated inflation and output gap, instead of the realized values. Orphanides (1999) discusses the difficulties and perils of implementing a Taylor-type rule in real time.

Strictly speaking, the output component in Taylor's original rule involves the percent deviations of output from a smooth (linear) trend. In the model considered here, such deviations are permanent because output has a unit root in equilibrium, implying that there is no deterministic trend or steady-state value that output reverts to. As a consequence, a rule of that sort would prevent the economy from adjusting to its long-run equilibrium path. In contrast, the output gap $\{x_t\}$ follows a stationary process, so that its inclusion in the rule will not lead to permanent deviations of the interest rate from its natural level.

[44] Given x_t , the equilibrium process for output and employment can be determined by using the fact that $y_t = x_t + y_t^*$, $c_t = x_t + c_t^*$, and $n_t = x_t + n_t^*$, with y_t^*, c_t^*, and n_t^* given by (2.11), (2.12), and (2.13).

Notice that in the particular case that both a_t and g_t follow a pure random walk ($\rho_a = 0$, $\rho_g = 1$), we have $x_t = \pi_t = 0$, all t; that is, the Taylor rule supports the optimal allocation. The reason is simple: The Taylor rule will implement the efficient allocation only if the latter is supported by a constant (natural) real rate, as is the case under the random walk assumption.

For more general driving processes, a monetary authority following a rule of the form of (5.7) could minimize the deviations from the optimal path by choosing sufficiently large values of ϕ_π and/or ϕ_x.[45] A Taylor rule with very high inflation or output gap coefficients, however, would potentially lead to huge instrument instability: Any small deviation of inflation or the output gap from zero (perhaps resulting from small measurement errors or imperfect credibility) would imply infinite changes in the rate.[46]

If $\rho_a \in (0, 1)$ and/or $\rho_g \in (0, 1)$, however, no finite values of the coefficients in rule (5.7) will replicate the optimal responses of output and inflation. The reason is straightforward: supporting the optimal response requires that prices remain stable and that the real and – given zero inflation – nominal rates change according to (5.3) in response to technology and fiscal shocks. However, a Taylor rule will not generate a change in the nominal rate unless a deviation from the optimal response arises in the form of a nonzero inflation or output gap.

To get a sense of the quantitative effects on macroeconomic stability and welfare of having a central bank follow a simple Taylor rule (instead of the optimal one), I have computed the standard deviations of inflation and the output gap predicted by a calibrated version of the sticky price model. I set $\phi_\pi = 1.5$ and $\phi_x = 0.5$, as in Taylor's original empirical rule. As just argued, avoiding the replication of the efficient allocation requires some departure from the random walk assumption for the driving variables. I set $\rho_a = 0.25$ and $\sigma_a = 0.01$, which roughly correspond to the first-order serial correlation of U.S. GDP growth, and to the standard deviation of the Solow residual innovations, respectively. To calibrate ρ_g and σ_g, I fit an AR(1) model to $g_t = -\log(1 - \tau_t)$, using the ratio of government purchases to GDP in the United States as the empirical counterpart to τ_t. That procedure yields values of $\rho_g = 0.95$ and $\sigma_g = 0.0036$. The remaining parameters are set at their baseline values, but results are also reported for two parameter variations: a calibration with higher risk aversion ($\sigma = 5$), as well as one with weaker nominal rigidities ($\theta = 0.5$).

The first set of columns in Table 5.1 displays some summary statistics for the resulting equilibrium. In the baseline case, the volatility of both inflation and the output gap is extremely small, independent of the type shock; as a result the implied welfare losses associated with the deviations from the efficient allocation are tiny, amounting to less than one hundredth of a percent of steady-state

[45] Formally, this is the case because Ω (and, hence, $\mathbf{B}_T$) converges to zero as ϕ_π or ϕ_x approaches infinity.

[46] Furthermore, the lack of credibility of such a policy might be more than warranted because it would easily hit the zero bound on the nominal rate.

Table 5.1. *Properties of three simple monetary policy rules*

	Taylor Rule			Money Growth Peg			Interest Rate Peg		
	Supply	Shocks Demand	Both	Supply	Shocks Demand	Both	Supply	Shocks Demand	Both
Baseline									
$\sigma(\pi)$	0.15	0.15	0.21	2.29	0.18	2.30	2.29	0.28	2.31
$\sigma(x)$	0.16	0.01	0.16	0.98	0.11	0.98	0.98	0.16	0.99
% Welfare loss	0.001	0.001	0.002	0.221	0.001	0.222	0.22	0.003	0.22
$\sigma = 5$									
$\sigma(\pi)$	0.23	0.30	0.38	0.70	0.30	0.76	3.42	0.39	3.44
$\sigma(x)$	0.08	0.007	0.08	0.08	0.05	0.10	0.40	0.06	0.41
% Welfare loss	0.002	0.003	0.006	0.020	0.003	0.023	0.474	0.006	0.48
$\theta = 0.5$									
$\sigma(\pi)$	0.47	0.20	0.51	3.18	0.31	3.20	3.18	0.46	3.22
$\sigma(x)$	0.08	0.002	0.08	0.38	0.05	0.39	0.38	0.07	0.39
% Welfare loss	0.0016	0.0003	0.002	0.070	0.001	0.071	0.070	0.001	0.072

consumption. Things do not change dramatically when we increase the degree of risk aversion ($\sigma = 5$) or lower the degree of price stickiness ($\theta = 5$); most noticeably, however, the volatility of inflation increases in both cases while that of the output gap decreases. Either way, the utility losses resulting from following a Taylor rule instead of the optimal one remain very small.

5.3.2. Money Growth Peg

Next I study the consequences of having the monetary authority maintain a constant rate of growth for the money supply, in the face of both supply and demand shocks. Without loss of generality, and for consistency with the steady state with zero inflation and no secular output growth, I assume $\Delta m_t = 0$, for all t. After defining $mpy_t = m_t - p_t - \overline{y}_t$, we can represent the equilibrium dynamics under that specification of monetary policy by means of the following stationary system:

$$\begin{bmatrix} 1+\frac{1}{\sigma\eta} & 0 & 0 \\ -\kappa & 1 & 0 \\ 0 & -1 & 1 \end{bmatrix} \begin{bmatrix} x_t \\ \pi_t \\ mpy_{t-1} \end{bmatrix} = \begin{bmatrix} 1 & \frac{1}{\sigma} & \frac{1}{\sigma\eta} \\ 0 & \beta & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} E_t\{x_{t+1}\} \\ E_t\{\pi_{t+1}\} \\ mpy_t \end{bmatrix}$$
$$+ \begin{bmatrix} \psi_a\rho_a & (1-\psi_g)(1-\rho_g) & 0 \\ 0 & 0 & 0 \\ \psi_a & \psi_g & -\psi_g \end{bmatrix} \begin{bmatrix} \Delta a_t \\ g_t \\ g_{t-1} \end{bmatrix}.$$

The second set of columns in Table 5.1 summarizes some of the statistical properties of the resulting equilibrium. The volatility of both inflation and the output gap resulting from technology shocks is substantially greater than that observed under a Taylor rule, whereas the difference is less pronounced for demand shocks. The welfare losses are no longer negligible, amounting to one quarter of a percent of steady-state consumption when both shocks are operative.

5.3.3. Interest Rate Peg

The third rule considered here consists of pegging the nominal interest rate at a level $r_t = \rho$, that is, the level consistent with a zero steady-state inflation. If interpreted as a rule followed mechanically by the central bank, such a specification of monetary policy is only a particular case of a simple Taylor rule with $\phi_\pi = \phi_x = 0$. But, as argued herein, such a configuration of parameters renders the equilibrium indeterminate. What rule would support a constant nominal rate while guaranteeing determinacy? Consider the following candidate:

$$r_t = \rho + \phi_r(\pi_t + \sigma\,\Delta c_t), \tag{5.8}$$

where $\phi_r > 1$. Combining the previous rule with (2.4) yields the difference equation

$$\pi_t + \sigma\,\Delta c_t = \phi_r^{-1} E_t\{\pi_{t+1} + \sigma\,\Delta c_{t+1}\},$$

whose only stationary solution satisfies $\pi_t = -\sigma \Delta c_t$, which in turn implies a constant nominal rate $r_t = \rho$, for all t. Rule (5.8) can be viewed as a modified Taylor rule, with consumption growth replacing the output gap. In addition, the inflation and consumption growth coefficients must satisfy a certain proportionality condition: the size of the coefficient on consumption must be exactly σ times that of inflation (which in turn must be greater than one). Notice also that the assumption of log utility ($\sigma = 1$) combined with the absence of fiscal shocks ($\Delta c_t = \Delta y_t$) implies a constant nominal GDP; in that case an interest rate peg is equivalent to strict nominal income targeting, and both can be implemented by (5.8).

Using the fact that $\pi_t = -\sigma(\Delta x_t + \Delta c_t^*)$, combined with (2.19) and (2.12), we can represent the equilibrium dynamics by the following system:

$$\begin{bmatrix} \pi_t \\ x_t \\ x_{t-1} \end{bmatrix} = \begin{bmatrix} \beta & 0 & \kappa \\ 0 & 0 & 1 \\ 0 & -\beta & 1+\beta+\frac{\kappa}{\sigma} \end{bmatrix} \begin{bmatrix} E_t\{\pi_{t+1}\} \\ E_t\{x_{t+1}\} \\ x_t \end{bmatrix} + \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ b_1 & b_2 & b_3 \end{bmatrix} \begin{bmatrix} \Delta a_t \\ g_t \\ g_{t-1} \end{bmatrix},$$

where $b_1 = \psi_a(1 - \beta\rho_a)$, $b_2 = -(1 - \psi_g)[1 + \beta(1 - \rho_a)]$, and $b_3 = (1 - \psi_g)$.

The third section in Table 5.1 ($\theta = 5$) reports the volatility of output and inflation as well as the welfare losses under an interest rate peg. Notice that, under the assumption of $\sigma = 1$, and conditional on technology shocks being the only source of fluctuations, the statistics shown match exactly those under a money growth peg. This is no coincidence: in that case, a constant money growth implies a constant nominal rate, and vice versa; accordingly, the resulting allocations are identical and, as just discussed, they are characterized by sizable fluctuations in both the output gap and inflation. In the high-risk-aversion case ($\sigma = 5$), that equivalence no longer holds, and the deviations from the optimal allocation are more pronounced under the interest rate peg (the welfare losses in the latter case get close to half a percent of steady-state consumption). When demand shocks are the source of fluctuations, the performance of an interest rate peg is uniformly worse than a money growth peg, and far worse than a Taylor rule.

5.3.4. Discussion

The analysis just given has sought to illustrate, in the context of a basic sticky price model, some of the results on the properties of simple rules found in the literature. Some general lessons can be drawn.

First, supporting as an equilibrium the combination of prices and quantities that the central bank seeks to attain – in our case, the flexible price allocation with zero inflation – generally requires (possibly large) variations in *both* interest rates and monetary aggregates. Hence, simple policy rules that keep one of

those variables constant is likely to cause significant deviations from the desired outcome.

Second, we expect that in the target allocation the pattern of response of interest rates and monetary aggregates will generally differ across shocks. Hence, the fact that a simple rule provides a good approximation to the optimal policy conditional on a certain source of fluctuations does not guarantee a good performance if other shocks become dominant.

Third, a simple interest rate rule à la Taylor does a remarkably good job at approximating the outcome of the optimal policy in a calibrated version of the basic sticky price model. A similar result can be found in a variety of papers that use related, but not identical, frameworks. Furthermore, several authors have emphasized the robustness of the Taylor rule across a variety of models, relative to more complex policy rules (including those that are optimal for *some* model).[47]

A key feature of the simple Taylor rule considered here is that it makes the interest rate respond (in a stabilizing direction) to deviations from target in inflation and the output gap, that is, precisely the variables that enter the loss function used to evaluate alternative policies. That feature underlies, undoubtedly, the good performance of the rule. From that point of view, it is worth noting that the specification of a simple Taylor rule used here is likely to be more desirable than an alternative, more conventional one, in which the interest rate is assumed to respond to deviations of output from trend (instead of its deviation from its natural rate). As stressed in McCallum and Nelson (1999), Rotemberg and Woodford (1999), and Galí (2000a), a strong response to (detrended) output may be highly inefficient when shocks to fundamentals call for large changes in output. The output gap induced by such a policy will, in turn, lead to unnecessary fluctuations in inflation. It is not surprising, therefore, that a pure inflation targeting rule (i.e., one with little or no weight attached to output stabilization) may often do better than one that is also concerned about output stabilization.[48]

5.4. Optimal Monetary Policy in the Presence of an Output Gap–Inflation Trade-off

As discussed earlier, the baseline sticky price model analyzed herein embeds no tradeoff between inflation and output-gap stabilization: By following an appropriate policy, the monetary authority can attain the efficient allocation, that is, the one corresponding to an equilibrium in which both inflation and the output gap are constant.

[47] See, among others, Ireland (2000), Rudebusch and Svensson (1999), Levin, Wieland, and Williams (1999), Rotemberg and Woodford (1999), and Galí et al. (2000).

[48] Orphanides (1999) stresses an additional advantage of an inflation targeting rule: it avoids the risks associated with having the monetary authority respond to output-gap indicators ridden with large and persistent measurement error. According to Orphanides' analysis, that problem may have been at the root of the great inflation of the 1970s in the United States.

The lack of an inflation–output trade-off is viewed by many economists as an unappealing feature of the previous framework. That consideration has led some authors to amend the basic model in a way that adds some realism to the policy maker's problem, while preserving the tractability of the original model.[49] A simple, largely ad hoc, way to achieve that objective is to augment the inflation equation with a disturbance that generates a trade-off between inflation and the output gap. Formally,

$$\pi_t = \beta E_t\{\pi_{t+1}\} + \kappa x_t + u_t, \tag{5.9}$$

where u_t is a shock that implies a change in the equilibrium level of inflation consistent with the natural level of output. That shock is often referred to in the literature as a cost-push shock. For simplicity I assume that $\{u_t\}$ follows a white noise process with a zero mean and variance σ_u^2. The existence of a trade-off can be seen clearly by solving (5.9) forward, which yields

$$\pi_t = \kappa \sum_{k=0}^{\infty} \beta^k E_t\{x_{t+k}\} + u_t.$$

Thus, an adverse cost-push shock (a positive realization in u_t) necessarily leads to a rise in inflation, and/or a negative output gap (current or anticipated).

Consider, as a benchmark, the macroeconomic and welfare implications of a monetary policy that would fully accommodate the inflationary effects of a cost-push shock, by maintaining output at its natural level at all times. In that case we would have $x_t = 0$ and $\pi_t = u_t$ for all t. Using (5.2) we can derive an expression for the expected welfare loss (expressed as a percent of steady-state consumption) incurred under the fully accommodating policy:

$$\frac{\varepsilon\sigma_u^2}{2\lambda} \times 100.$$

If we normalize the standard deviation of the shock to be 1 percent ($\sigma_u = 0.01$), we see that, under the baseline calibration, that welfare loss (relative to the efficient allocation) amounts to 0.64 percent of steady-state consumption.

Of course, there is no reason why a central bank would want to accommodate fully a cost-push shock. Instead, it will want to respond to such a shock so that the resulting path of inflation and the output gap minimizes the utility loss. Formally, and given in (5.1), the central bank will seek to minimize

$$E_0\left\{\sum_{t=0}^{\infty} \beta^t \left[\kappa x_t^2 + \varepsilon\pi_t^2\right]\right\}, \tag{5.10}$$

subject to the sequence of budget constraints (5.9). Given the resulting optimal path for $\{x_t, \pi_t\}$, we can use (2.20) to derive the interest rate policy that would support such a path.

The form of the solution to the problem just given depends critically on the assumptions we make regarding the central bank's ability to commit to future

[49] See, e.g., Clarida et al. (1999).

policy actions. In the paragraphs that follow, I describe the optimal policy and its macroeconomic implications under two alternative, polar assumptions: discretion vs. commitment.

Before turning to the formal analysis, I find it important to stress the connection between the results presented here and the early literature on credibility and gains from commitment. The latter, exemplified by the work of Kydland and Prescott (1977), Barro and Gordon (1983), and Rogoff (1985), brought to light the risk of a persistent inflation bias arising from the central bank's inability to commit to a low inflation policy. In their framework, the ultimate source of that bias could be found in the central bank's desire to push output above its natural (flexible price) level, presumably because the latter is inefficiently low.[50] Without such a bias, an efficient outcome characterized by zero output and zero inflation can be attained under discretion. In other words, in the absence of an inflation bias there would not be any gains from commitment.

In the optimizing sticky price model considered here, the previous result no longer holds. Instead, and as shown in Clarida et al. (1999) and Woodford (1999b), *even in the absence of an inflation bias*, there are potential welfare gains associated with the central bank's ability to commit credibly to a systematic pattern of response to shocks that generate a trade-off between output and inflation. Next I show that result using the baseline NK model.

5.4.1. Optimal Discretionary Policy

Suppose that the monetary authority cannot credibly commit to any future policy actions. Because it is unable to influence current expectations on future output and inflation, it has to take those expectations as given. Accordingly, each period it will choose (x_t, π_t) in order to minimize

$$\kappa x_t^2 + \varepsilon \pi_t^2$$

subject to

$$\pi_t = \kappa x_t + \theta_t,$$

where $\theta_t = \beta E_t\{\pi_{t+1}\} + u_t$ is taken as given by the central bank. The solution to this problem must satisfy

$$x_t = -\varepsilon \pi_t. \tag{5.11}$$

Substituting (5.11) into (5.9), and solving the resulting difference equation forward, yields

$$\pi_t = \left(\frac{1}{1+\kappa\varepsilon}\right) u_t, \tag{5.12}$$

$$x_t = -\left(\frac{\varepsilon}{1+\kappa\varepsilon}\right) u_t. \tag{5.13}$$

[50] In that case, the terms involving x_t^2 in the loss function are replaced with $(x_t - x^*)^2$, where $x^* > 0$ denoted the output-gap target.

Hence, in response to an adverse cost-push shock, the central bank finds it optimal to engineer a temporary reduction in output, thus dampening the effect of the shock on inflation. The incentive to "split" the effects of the shock between output and inflation is a consequence of the convexity of the loss function in those variables.

The expected welfare loss under the optimal discretionary policy is given by

$$\frac{\varepsilon\sigma_u^2}{2\lambda(1+\kappa\varepsilon)} \times 100,$$

which is always lower than the loss under the fully accommodating policy.

Under the baseline calibration, the standard deviations of the output gap and (annualized) inflation when the central bank pursues the optimal discretionary policy are 3.80 percent and 1.38 percent, respectively. The implied welfare loss amounts now to 0.22 percent of steady-state consumption.

Using (2.20), we see that the implied equilibrium interest rate under the optimal discretionary policy is given by

$$r_t = \overline{rr}_t + \left(\frac{\sigma\varepsilon}{1+\kappa\varepsilon}\right)u_t.$$

Notice that a simple interest rate rule that would support (5.12) and (5.13) is given by

$$r_t = \overline{rr}_t + \sigma\varepsilon\pi_t. \tag{5.14}$$

As discussed earlier, the previous rule will guarantee that the desired allocation obtains if and only if $\sigma\varepsilon > 1$, a condition that is satisfied for plausible values of those parameters. Alternatively, a rule of the form

$$r_t = \overline{rr}_t + \left(\frac{\sigma\varepsilon - \phi_\pi}{1+\kappa\varepsilon}\right)u_t + \phi_\pi\pi_t,$$

with $\phi_\pi > 1$ will always guarantee uniqueness of the equilibrium.

5.4.2. *Optimal Policy With Commitment*

Suppose that the monetary authority can choose, once and for all, a state-contingent policy $\{x(z^t), \pi(z^t)\}_{t=0}^{\infty}$, where z^t denotes the history of shocks up to period t, and assume it sticks to its plan. The equilibrium dynamics for the output gap and inflation under the optimal policy with commitment can be shown to satisfy the optimality condition[51]

$$x_t = -\varepsilon(p_t - p^*), \tag{5.15}$$

[51] See Woodford (1999b) and Clarida et al. (1999) for a derivation. The latter paper also analyzes the (simpler) case of commitment under the constraint that both the output gap and inflation are a function of the current state only, not of its entire history.

for $t = 0, 1, 2, \ldots$, where p^* can be interpreted as a price level target, which corresponds to the (log) price level in the period before the monetary authority chooses (once and for all) its optimal plan (i.e., $p^* = p_{-1}$).

Combining (5.15) with (5.9), we can derive a stochastic difference equation for the output gap implied by the optimal policy:

$$x_t = ax_{t-1} + a\beta E_t\{x_{t+1}\} - a\varepsilon u_t,$$

for $t = 0, 1, 2, \ldots$, where $a = 1/(1 + \beta + \kappa\varepsilon)$. The (nonexplosive) solution to the previous difference equation is given by

$$x_t = \delta x_{t-1} - \varepsilon\psi_u u_t, \tag{5.16}$$

where $\delta = [(1 - \sqrt{1 - 4\beta a^2})/2a\beta] \in (0, 1)$ and $\psi_u = a(1 + \beta\delta^2)$. We can then use (5.15) to derive the equilibrium process for $\widetilde{p}_t = p_t - p^*$, the deviation of the price level from target:

$$\widetilde{p}_t = \delta\widetilde{p}_{t-1} + \psi_u u_t. \tag{5.17}$$

Thus, we see that the optimal policy with commitment implies a stationary process for the price level; that is, the deviations of the price level from the target level p^* are only transitory, with any inflation resulting from a cost-push shock being eventually followed by deflation.

The pattern of responses for (annualized) inflation and the output gap under commitment is depicted graphically in Figure 5.7, which also displays the responses under the optimal discretionary policy. The figure illustrates two aspects of the differences in outcome that are worth pointing out. First, one can show that the possibility of commitment improves the terms of the output–inflation trade-off facing the policy maker. This is illustrated in the figure, where we see that the increase in inflation resulting from a unit cost-push shock is smaller under commitment than under discretion, even though the associated decline in the output gap is also smaller. That result has a simple explanation, directly related to the forward-looking nature of inflation: the response of inflation to a cost-push shock depends on the anticipated path for the output gap. Under discretion, the output gap returns to zero once the shock dies out. Accordingly, the initial response of inflation is proportional to the initial decline in the output gap, given the shock. By way of contrast, under commitment the output gap remains negative well after the direct effects of the shock have vanished, and returns to its initial level only asymptotically; the anticipation of that low level of economic activity in the future has, in itself, a dampening effect on inflation, thus explaining the smaller rise of inflation under commitment, despite the smaller decline in output.

Second, it is clear that the joint pattern of output and inflation under commitment in the periods following the shock is time inconsistent. Because both inflation and the output gap take on negative values (and thus generate a welfare loss), it would be optimal (as well as feasible) for a central bank that did not

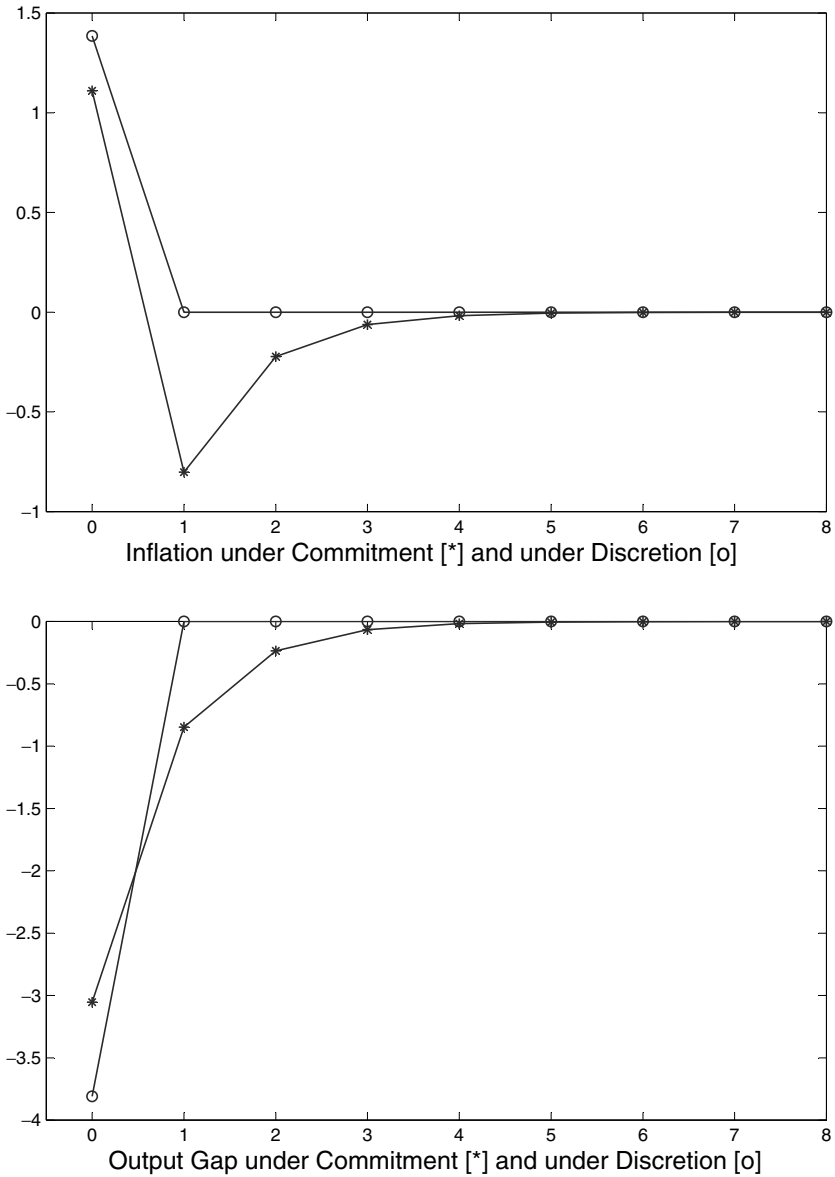


Figure 5.7. Commitment vs. discretion.

feel restrained by its earlier promises to pursue a more expansionary policy and to restore the efficient allocation.

Because the optimal discretionary policy falls within the range of feasible policies available to a policy maker with access to a commitment technology, it is not surprising that the welfare losses associated with the optimal policy

with commitment are smaller than in the discretion case. Thus, for instance, under our baseline calibration, the welfare loss represents 0.17 percent of steady-state consumption (compared with 0.22 percent in the discretionary case). Underlying that loss in welfare in the numerical example, we have standard deviations of the output gap and (annualized) inflation given by 3.17 percent and 1.39 percent, respectively (compared with 3.80 percent and 1.38 percent under discretion).

6. STAGGERED WAGE SETTING

The analysis up to this point has been based on a stylized model in which imperfect competition and sticky prices in the goods market coexist with a labor market characterized by perfect competition and flexible wages. Up until recently, that was the framework that seemed to be favored in macroeconomic analysis. One possible reason for that choice may be that an alternative framework that emphasized wage stickiness (and perfectly competitive good markets) would tend to predict, counterfactually, a procyclical behavior of real wages, at least if demand shocks were the dominant source of fluctuations.

Yet, several authors have recently studied the consequences of introducing nominal rigidities (with or without price rigidities) in a dynamic general equilibrium framework. Here I lay out an extension of the baseline sticky price model that incorporates staggered wage setting.[52]

6.1. A Simple Model of Staggered Wage and Price Setting

Consider a continuum of labor types, indexed by $j \in [0, 1]$. The effective labor input in the production process of a typical firm is given by a CES function of the quantities of different types of labor hired. Thus, for firm i,

$$N_{it} = \left[\int_0^1 N_{it}(j)^{(\varepsilon_w - 1)/\varepsilon_w} dj \right]^{\varepsilon_w/(\varepsilon_w - 1)},$$

where $\varepsilon_w > 1$ is the elasticity of substitution between labor types in production.

The quantity of labor of type j demanded by firm i is given by

$$N_{it}(j) = \left[\frac{W_t(j)}{W_t} \right]^{-\varepsilon_w} N_{it},$$

where $W_t(j)$ is the cost of hiring one unit of type j labor and $W_t = [\int_0^1 W_t(j)^{1-\varepsilon_w} dj]^{1/(1-\varepsilon_w)}$ is an aggregate wage index.

[52] The work of Blanchard and Kiyotaki (1987) contains an early analysis of a static model with both wage and price stickiness. The model developed here is a simplified version of the one analyzed in Erceg, Henderson, and Levin (2000). Kim (2000) develops a model in a similar spirit, but in which the stickiness of wages and prices results from the existence of adjustment costs. Erceg (1997) and Huang and Liu (1998) analyze the effects of exogenous monetary policy shocks in a model with wage contracts à la Taylor.

Each household specializes in supplying a differentiated type of labor. Total demand for labor of type j, $N_t(j) = \int_0^1 N_{it}(j)di$, is given by

$$N_t(j) = \left[\frac{W_t(j)}{W_t}\right]^{-\varepsilon_w} N_t,$$

where $N_t = \int_0^1 N_{it}di$ denotes aggregate labor input.

If all households could set their wage optimally every period, they would do so according to the log-linear rule $w_t = \mu_w + \text{mrs}_t + p_t$, where $\mu_w = \log[\varepsilon_w/(\varepsilon_w - 1)]$ represents the desired markup of the real wage over the marginal rate of substitution between consumption and leisure $\text{mrs}_t = \sigma c_t + \varphi n_t$.

Staggered wage setting is introduced by assuming that each household faces a probability θ_w of having to keep the wage for its labor type unchanged in any given period. Let w_t^* denote the (log) wage set by households adjusting wages in period t. The evolution of the aggregate wage level over time can be approximated by the log-linear difference equation

$$w_t = \theta_w w_{t-1} + (1 - \theta_w)w_t^*. \tag{6.1}$$

One can show that optimizing households will set their wage according to the (approximate) log-linear rule

$$w_t^* = \mu_w + (1 - \beta\theta_w)\sum_{k=0}^{\infty}(\beta\theta_w)^k E_t\{\text{mrs}_{t,t+k} + p_{t+k}\}, \tag{6.2}$$

where $\text{mrs}_{t,t+k} = \sigma c_{t+k} + \varphi n_{t,t+k}$, with $n_{t,t+k}$ being the quantity of labor supplied in period $t + k$ by a household whose wage was last reset in period t.[53] The intuition behind the previous wage setting rule is analogous to that for price setting: Households (workers) will set the nominal wage so that a weighted average of the expected wage markup over the duration of the contract matches the frictionless optimal markup μ_w.

Using the fact that

$$n_{t,t+k} = -\varepsilon_w(w_t^* - w_{t+k}) + n_t,$$

we can rewrite (6.2) in terms of aggregate variables as follows:

$$w_t^* - w_t = -\left(\frac{1 - \beta\theta_w}{1 + \varphi\varepsilon_w}\right)\sum_{k=0}^{\infty}(\beta\theta_w)^k E_t\{\hat{\mu}_{t+k}^w\} + \sum_{k=1}^{\infty}(\beta\theta_w)^k E_t\{\pi_{t+k}^w\}, \tag{6.3}$$

where $\hat{\mu}_t^w = (w_t - p_t) - \text{mrs}_t - \mu^w$ can be interpreted as the percent deviation between the average wage markup and its frictionless level.

[53] Notice that the level of consumption is the same across households, as a result of risk sharing.

Combining (6.1) and (6.3), we obtain the wage inflation equation

$$\pi_t^w = \beta E_t\{\pi_{t+1}^w\} - \lambda_w \hat{\mu}_t^w, \tag{6.4}$$

where $\lambda_w = [(1-\beta\theta_w)(1-\theta_w)]/[\theta_w(1+\varphi\varepsilon_w)]$. Also notice the following relationship between the wage markup and the real marginal cost:

$$\begin{aligned} \mathrm{mc}_t &= (w_t - p_t) - a_t \\ &= \mu_t^w + \mathrm{mrs}_t - a_t \\ &= \mu_t^w + (\sigma+\varphi)y_t - (1+\varphi)a_t - \sigma g_t. \end{aligned}$$

It follows that $\widehat{\mathrm{mc}}_t = (\sigma+\varphi)x_t + \hat{\mu}_t^w$, where x_t denotes the (log) deviation of output from its level in the absence of both price and wage rigidities.[54]

One can combine the previous results with (2.17) (whose derivation was independent of the presence or not of wage rigidities) to obtain a version of the NPC for an economy with staggered wage setting:

$$\pi_t = \beta E_t\{\pi_{t+1}\} + \kappa x_t + \lambda \hat{\mu}_t^w. \tag{6.5}$$

Interestingly, (6.5) seems to provide a theoretical justification for the inclusion of the cost-push disturbance in (5.9), as a way to generate a trade-off between output and inflation. Yet, an important difference remains: the disturbance thus generated cannot be assumed to be exogenous, because it will generally depend on preference and technology parameters, as well as on the underlying disturbances.[55]

6.1.1. A Particular Case: Sticky Wages and Flexible Prices

Consider the polar case with sticky wages but fully flexible prices. In that case, all firms face identical marginal costs and charge identical prices, consistent with the desired markup (i.e., $\hat{\mu}_t^p = 0$). As a result, there is no relative price distortion and the variance of price inflation should no longer be a central bank's concern. In that case, it is possible for the central bank to replicate the flexible price and wage allocation by having $\hat{\mu}_t^w = \pi_t^w = x_t = 0$, for all t. It follows that price inflation under the optimal policy will be given by $\pi_t = -\Delta a_t$, for all t. The prescription of full price stabilization is no longer desirable in that environment; the monetary authority should seek to stabilize wages instead.

The nominal rate that will support that efficient allocation will have to satisfy

$$\begin{aligned} r_t &= \overline{rr}_t + E_t\{\pi_{t+1}\} \\ &= \rho + \frac{(\sigma-1)\varphi\rho_a}{\sigma+\varphi}\Delta a_t + \sigma(1-\psi_g)(1-\rho_g)g_t. \end{aligned}$$

[54] Equivalently, we see that $x_t = -[(\hat{\mu}_t^p + \hat{\mu}_t^w)/(\sigma+\varphi)]$; that is, the output gap is proportional to the sum of the deviations of the price and wage markups from their steady-state levels. Because $\hat{\mu}_t^p + \hat{\mu}_t^w = \mathrm{mrs}_t - a_t$, it follows that the output gap is proportional to the wedge between the marginal rate of substitution between consumption and labor and the corresponding marginal rate of transformation, and hence it is a measure of the (uncorrected) aggregate distortions in the economy.

[55] See Erceg et al. (2000).

6.2. Optimal Monetary Policy When Both Wages and Prices Are Sticky

Can the allocation associated with flexible prices and wages be restored when both prices and wages are sticky? The answer is no, as shown in Erceg, Henderson, and Levin (2000; henceforth EHL). The reason is straightforward: replicating that allocation requires $\hat{\mu}_t^p = \hat{\mu}_t^w = 0$, for all t. It would then follow from (2.17) and (6.4) that $\pi_t^w = \pi_t = 0$, which in turn implies a constant real wage. However, that is inconsistent with the requirement that the real wage adjusts on a one-to-one basis with the marginal product of labor and the marginal rate of substitution if the real marginal cost and the wage markup are to remain constant. To the extent that the equilibrium in the model with flexible wages and prices involves fluctuations in either variable (as will generally be the case), the efficient allocation will not be attainable.

Given the existence of a trade-off among stabilization of the output gap, price inflation, and wage inflation, what is the appropriate course of action for a policy maker seeking to maximize consumers' welfare? EHL have derived the loss function for an economy with wage and price stickiness, thus generalizing the analysis of Rotemberg and Woodford. The resulting loss function, expressed as a fraction of steady-state consumption, is given by

$$-\frac{1}{2}\left[(\sigma+\varphi)\text{var}(x_t)+\frac{\varepsilon}{\lambda}\,\text{var}(\pi_t)+\frac{\varepsilon_w}{\lambda_w}\,\text{var}\left(\pi_t^w\right)\right].$$

Notice that λ and λ_w are decreasing in the degree of price and wage rigidities (respectively), as parameterized by θ_p and θ_w. Hence, the monetary authority will attach a relatively greater weight to price (wage) inflation stabilization the stronger (weaker) price rigidities are relative to wage rigidities.[56] The baseline sticky price model analyzed in the previous sections corresponds to the limiting case $\lambda_w \to +\infty$, so that wage inflation stabilization stops being a concern (a symmetric result holds for the flexible price case). Notice that the welfare loss associated with a given level of price or wage inflation is proportional to the elasticities of substitution among different goods and different types of labor, respectively. That result reflects the fact that the degree of substitutability will enhance the (inefficient) dispersion in output and employment levels generated by the staggering of prices and/or wages in an environment with nonzero inflation.

EHL (2000) use numerical methods to derive the optimal policy rule in an economy similar to the one just described, and they determine the implied volatility of the output gap, price inflation, and wage inflation, as well as the associated welfare losses. Among other results, they show that when prices are more (less) rigid than wages, the optimal policy requires that wages (prices) account for a relatively larger share of the real wage adjustment; as a consequence, the rate of inflation of the more flexible variable ends up displaying

[56] Benigno (1999) obtains a related result when analyzing the optimal policy in a monetary union: The central bank should put more weight to stabilization of the rate of inflation in the economy facing stronger nominal rigidities.

higher volatility (in a way consistent with its smaller weight in the loss function). EHL also show that, for a variety of calibrations, simple rules that put a lot of weight on wage inflation and/or output-gap stabilization perform nearly as well as the optimal one.

7. CONCLUDING REMARKS

This paper has surveyed a number of results generated by recent research on monetary policy in dynamic optimizing models with nominal rigidities. In my opinion, that research program has yielded several new insights, as well as a number of results that one may view as surprising, regarding the linkages among monetary policy, inflation, and the business cycle. In other words, and contrary to what some economists might have predicted, the effort to integrate Keynesian-type elements into a dynamic general equilibrium framework has gone beyond "providing rigorous microfoundations" to some preexisting, though largely ad hoc, framework. Furthermore, that research program is making significant progress toward the development of a *standard* framework that can be used meaningfully for the purpose of evaluating alternative monetary policies. Perhaps the clearest proof of that potential lies in the renewed interest shown by many central banks in the recent academic research on monetary economics.

ACKNOWLEDGEMENTS

This paper was prepared for an invited session at the World Congress of the Econometric Society, August, 2000. I thank Chris Sims for a useful discussion at the conference. Several sections of the paper draw on previous work of mine with several coauthors, including Richard Clarida, Mark Gertler, David López-Salido, and Javier Vallés. Financial support from the National Science Foundation, the C.V. Starr Center for Applied Economics, and CREI is gratefully acknowledged. Correspondence: CREI, Ramon Trias Fargas 25, 08005 Barcelona (Spain). E-mail: jordi.gali@econ.upf.es. Web page: www.econ.upf.es/∼gali.

References

Andersen, T. M. (1998), "Persistency in Sticky Price Models," *European Economic Review*, 42, 593–603.

Andrés, J., D. López-Salido, and J. Vallés (1999), "Intertemporal Substitution and the Liquidity Effect in a Sticky Price Model," mimeo, Bank of Spain.

Basu, S., J. Fernald, and M. Kimball (1998), "Are Technology Improvements Contractionary?" mimeo, New York University.

Benigno, P. (1999), "Optimal Monetary Policy in a Currency Area," mimeo, Princeton University.

Bergin, P. R. and R. C. Feenstra (2000), "Staggered Price Setting, Translog Preferences, and Endogenous Persistence," *Journal of Monetary Economics*, 45, 657–680.

Bernanke, B. S., and I. Mihov (1998), "The Liquidity Effect and Long Run Neutrality," *Carnegie–Rochester Series on Public Policy,* 49, 149–194.

Bernanke, B. S. and M. Woodford (1997), "Inflation Forecasts and Monetary Policy," *Journal of Money, Credit and Banking*, 24, 653–684.

Blanchard, O. J. (1997), "Comment on The New Neoclassical Synthesis and the Role of Monetary Policy," *NBER Macroeconomics Annual*, 1997, 289–293.

Blanchard, O. J. and C. M. Kahn (1980), "The Solution of Linear Difference Models Under Rational Expectations," *Econometrica,* 48(5), 1305–1312.

Blanchard, O. J. and N. Kiyotaki (1987), "Monopolistic Competition and the Effects of Aggregate Demand," *American Economic Review*, 77, 647–666.

Blinder, A. S. (1998), *Central Banking in Theory and Practice*. Cambridge, MA: MIT Press.

Bullard, J. and K. Mitra (1999), "Learning About Monetary Policy Rules," *Journal of Monetary Economics*, forthcoming.

Calvo, G. (1983), "Staggered Prices in a Utility Maximizing Framework," *Journal of Monetary Economics*, 12, 383–398.

Chadha, B., P. Masson, and G. meredith (1992) "Models of Inflation and the Costs of Disinflation 39(2)," *IMF Staff Papers*, 395–431.

Chari, V. V., P. J. Kehoe, and E. R. McGrattan (2000), "Sticky Price Models of the Business Cycle: Can the Contract Multiplier Solve the Persistence Problem?" *Econometrica*, 68(5), 1151–1180.

Chari, V. V., P. J. Kehoe, and E. R. McGrattan (1996), "Sticky Price Models of the Business Cycle: Can the Contract Multiplier Solve the Persistence Problem?," Research Department Staff Report 217, Federal Reserve Bank of Minneapolis.

Christiano, L. J., M. Eichenbaum, and C. L. Evans (1997), "Sticky Price and Limited Participation Models: A Comparison," *European Economic Review*, 41(6), 1201–1249.

Christiano, L. J., M. Eichenbaum, and C. L. Evans (1998), "Modeling Money," NBER Working Paper 6371.

Christiano, L. J., M. Eichenbaum, and C. L. Evans (1999), "Monetary Policy Shocks: What Have We Learned and to What End?," in *Handbook of Macroeconomics*, Vol. 1A (ed. by J. B. Taylor and M. Woodford), Amsterdam: North-Holland, 65–148.

Clarida, R., J. Galí, and M. Gertler (1998), "Monetary Policy Rules in Practice: Some International Evidence," *European Economic Review*, 42, 1033–1067.

Clarida, R. J. Galí, and M. Gertler (1999), "The Science of Monetary Policy: A New Keynesian Perspective," *Journal of Economic Literature*, 37(4), 1661–1707.

Clarida, R., J. Galí, and M. Gertler (2000), "Monetary Policy Rules and Macroeconomic Stability: Evidence and Some Theory," *Quarterly Journal of Economics*, 115(1), 147–180.

Cooley, T. F. and G. D. Hansen (1989), "Inflation Tax in a Real Business Cycle Model," *American Economic Review*, 79, 733–748.

Correia, I. and P. Teles (1999), "The Optimal Inflation Tax," *Review of Economic Dynamics*, 2(2), 325–346.

Dotsey, M. (1999), "Structure from Shocks," mimeo, Federal Reserve Bank of Richmond.

Dotsey, M., R. G. King, and A. L. Wolman (1999), "State Dependent Pricing and the General Equilibrium Dynamics of Money and Output," *Quarterly Journal of Economics*, 114(2), 655–690.

Erceg, C. J. (1997), "Nominal Wage Rigidities and the Propagation of Monetary Disturbances," mimeo, Federal Reserve Board.

Erceg, J., D. W. Henderson, and A. T. Levin (2000), "Optimal Monetary Policy with Staggered Wage and Price Contracts," *Journal of Monetary Economics*, 46(2), 281–314

Fuhrer, J. C. (1997), "The (Un)Importance of Forward-Looking Behavior in Price Setting," *Journal of Money, Credit and Banking*, 29, 338–350.

Fuhrer, J. C. and G. R. Moore (1995a), "Inflation Persistence," *Quarterly Journal of Economics*, 110, 127–159.

Galí, J. (1999), "Technology, Employment, and the Business Cycle: Do Technology Shocks Explain Aggregate Fluctuations?" *American Economic Review*, 89(1), 249–271.

Galí, J. and M. Gertler (1999), "Inflation Dynamics: A Structural Econometric Analysis," *Journal of Monetary Economics*, 44(2), 195–222.

Galí, J. M. Gertler, and D. López-Salido (2001), "European Inflation Dynamics," *European Economic Review*, 45(7), 1237–1270.

Galí, J., D. López-Salido, and J. Vallés (2000), "Technology Shocks and Monetary Policy: Assessing the Fed's Performance," *Journal of Monetary Economics*, forthcoming.

Galí, J., (2000a), "The Conduct of Monetary Policy in the Face of Technological Change: Theory and Postwar U.S. Evidence," mimeo. www.econ.upf.es/∼gali.

Galí, J. (2000b), "Targeting Inflation in an Economy with Staggered Price Setting," in *Inflation Targeting: Design, Performance, Challenges*, (ed. by N. Loayza and R. Soto), Santiago: Central Bank of Chile.

Goodfriend, M. and R. G. King (1997), "The New Neoclassical Synthesis and the Role of Monetary Policy," *NBER Macroeconomics Annual*, 231–283.

Hairault, J.-O., and F. Portier (1993), "Money, New Keynesian Macroeconomics, and the Business Cycle," *European Economic Review*, 37, 33–68.

Huang, K. X. D. and Z. Liu (1998), "Staggered Contracts and Business Cycle Persistence," mimeo. Minneapolis Fed Discussion Paper No. 27.

Ireland, P. N. (2000), "Interest Rates, Inflation, and Federal Reserve Policy since 1980," *Journal of Money, Credit, and Banking*, 32, 417–434.

Jeanne, O. (1998), "Generating Real Persistent Effects of Monetary Shocks: How Much Nominal Rigidity Do We Really Need?" *European Economic Review*, 42(6), 1009–1032.

Jensen, H. (1999), "Targeting Nominal Income Growth or Inflation," CEPR Discussion Paper 2341.

Judd, J. P. and G. Rudebusch (1998), "Taylor's Rule and the Fed: 1970–1997," *Economic Review*, 3, 3–16.

Khan, A., R. King, and A. Wolman (2000), "Optimal Monetary Policy," mimeo, Federal Reserve Bank of Richmond.

Kiley, M. (1997), "Staggered Price Setting and Real Rigidities," mimeo, Federal Reserve Board.

Kim, J. (2000), "Constructing and Estimating a Realistic Optimizing Model of Monetary Policy," *Journal of Monetary Economics*, 45, 329–359.

King, R. G. and A. L. Wolman (1996), "Inflation Targeting in a St. Louis Model of the 21st Century," *Federal Reserve Bank of St. Louis Review*, 78(3), NBER Working Paper 5507.

Kydland, F., and E. C. Prescott (1982) "Time to Build and Aggregate Fluctuations," *Econometrica*, 50, 1345–70.

Lane, Philip R. (2001): "The New Open Macroeconomics: A Survey," *Journal of International Economics*. 54(2), 235–266.

Leeper, E. M., C. Sims, and T. Zha (1996), "What Does Monetary Policy Do?" *Brookings Papers on Economic Activity*, 2, 1–78.

Levin, A. T., V. Wieland, and J. Williams (1999), "Robustness of Simple Monetary

Policy Rules under Model Uncertainty," in *Monetary Policy Rules*, (ed. by J. B. Taylor), Chicago: University of Chicago Press.

McCallum, B. and E. Nelson (1999), "Performance of Operational Policy Rules in an Estimated Semiclassic a Structural Model," in *Monetary Policy Rules*, (ed. by J. B. Taylor), Chicago: University of Chicago Press.

Nelson, E. (1998), "Sluggish Inflation and Optimizing Models of the Business Cycle," *Journal of Monetary Economics*, 42(2), 303–322.

Orphanides, A. (1999), "The Quest for Prosperity without Inflation," mimeo, Board of Governors.

Roberts, J. M. (1997), "Is Inflation Sticky?" *Journal of Monetary Economics*, 39, 173–196.

Rotemberg, J. (1996), "Prices, Output, and Hours: An Empirical Analysis Based on a Sticky Price Model," *Journal of Monetary Economics*, 37, 505–533.

Rotemberg, J. and M. Woodford (1997), "An Optimization-Based Econometric Framework for the Evaluation of Monetary Policy," *NBER Macroeconomics Annual*, 297–346.

Rotemberg, J. and M. Woodford (1999), "Interest Rate Rules in an Estimated Sticky Price Model," in *Monetary Policy Rules*, (ed. by J. B. Taylor), Chicago: University of Chicago Press.

Rotemberg, J. and M. Woodford (1999), "The Cyclical Behavior of Marginal Costs," in *Handbook of Macroeconomics*, Vol. 1B, (ed. by J. B. Taylor and M. Woodford), Amsterdam: North Holland, pp. 1051–1136.

Rudebusch, G. and L. Svensson (1999). "Policy Rules for Inflation Targeting" in *Monetary Policy Rules*, (ed. by J. B. Taylor), University of Chicago Press, 203–262.

Sbordone, A. (1998), "Prices and Unit Labor Costs: Testing Models of Pricing Behavior," Rutgers University.

Stock, J. and M. Watson (2000). "Business Cycle Fluctuations in U.S. Macroeconomic Time Series," in *Handbook of Macroeconomics*, (ed. J. B. Taylor and M. Woodford), 1A, 3–64.

Taylor, J. B. (1993), "Discretion Versus Policy Rules in Practice," *Carnegie–Rochester Conference Series on Public Policy*, 39, 195–214.

Taylor, J. B. (1998), "An Historical Analysis of Monetary Policy Rules," in *Monetary Policy Rules*, (ed. by J. B. Taylor), Chicago: University of Chicago Press.

Taylor, J. B. (1999), *Monetary Policy Rules*. Chicago: University of Chicago Press and NBER.

Vestin, D. (1999), "Price-Level Targeting vs. Inflation Targeting in a Forward Looking Model," mimeo, IIES.

Walsh, C. E. (1998), *Monetary Theory and Policy*, Chapters 2–4, Cambridge, MA: MIT Press.

Woodford, M. (1996), "Control of the Public Debt: A Requirement for Price Stability?," Working Paper 5684, NBER.

Woodford, M. (1999a), "Optimal Monetary Policy Inertia," NBER Working Paper 7261.

Woodford, M. (1999b), "How Should Monetary Policy Be Conducted in an Era of Price Stability?: A Comment," mimeo, Princeton University.

Woodford, M. (1999c), *Interest and Prices*. Chapter 6, mimeo, Princeton University.

Woodford, M. (2000), "Pitfalls of Forward-Looking Monetary Policy," mimeo, Princeton University.

Yun, T. (1996), "Nominal Price Rigidity, Money Supply Endogeneity, and Business Cycles," *Journal of Monetary Economics*, 37, 345–370.

Comments on Papers by Stefania Albanesi, V. V. Chari, and Lawrence J. Christiano and by Jordi Galí

Christopher A. Sims

The theme of these comments is that we need to remain vigilant against the possibility that "standard" modeling conventions in macroeconomics, originally introduced as experimental or tentative, start to be used unquestioningly, despite serious drawbacks. It is, in a sense, unfair to focus such comments on the papers presented in this session, because the conventions to be criticized are not at all special to these papers, but there must be some point at which we step back and consider where the conventions in our literature are headed, and a quinquennial World Congress seems as appropriate an occasion as any for doing this.

I consider the papers by Galí and by Albanesi, Chari, and Christiano (henceforth ACC) that appear in this volume; and also the empirical paper (Galí and Gertler, 1999) that forms the foundation for much of the paper in this volume that Jordi Galí presented.

1. OPTIMAL MONETARY AND FISCAL POLICY

Both papers discuss optimal monetary policy in the context of general equilibrium models. General equilibrium models, with their own internally generated, explicit measures of welfare, allow us to avoid postulating an ad hoc objective function for the policy authority, instead evaluating policy directly in terms of its welfare implications for private agents. Galí explicitly, and probably ACC also, see this as an improvement over approaches that use models with a less complete behavioral interpretation. Such models must either postulate plausible measures of policy effectiveness directly, or else present multidimensional measures of policy performance – variances and means of a variety of important variables, usually – and leave the reader to draw conclusions about ranking the results.

However, these papers only confirm that reliable conclusions about rankings of policies from general equilibrium macro models are, for the time being at least, completely unavailable. ACC focus most of their attention on two welfare effects of inflation – the decline in demand it generates for "cash goods," under their particular assumptions about timing in discrete time cash-in-advance style

models, and the increase in output toward the efficient level it can produce under sticky prices. Output tends to be below the efficient level in their models because of the existence of monopolistic competition. Even within their framework, their conclusions are more fragile and dubious than a casual reading of the paper might suggest.

Galí and the related literature he cites, in contrast, assume both these effects of monetary policy away at the start. The assumption that the costs of squeezing down real balances are small is justified by no more than an appeal to intuition. The assumption that there are no gains from pushing output above its noninflationary level, despite the same monopolistic competition assumption as in the ACC model, requires postulating that the government knows and sets exactly the efficient level of employment subsidy, financing it by lump sum taxes. The costs that are left are only those caused by inefficient, but mean zero, fluctuations around the efficient level of output.

That there is no overlap in the effects of monetary policy considered in these two papers, because they are representative of much other work in the recent literature, suggests that we have no professional consensus on how to set about this task. Furthermore, a good case can be made that each is considering potentially important effects, and thus also that each is ignoring potentially important effects.

It is common to dismiss as unimportant "shoe-leather costs" imposed by inducing people to economize on real balances, so in doing so Galí is far from alone. If we think of these as the costs imposed on us when inflation induces us to carry less cash in our wallets, it certainly seems justified to treat them as small. From this point of view, the fact that cash-in-advance and limited-participation models are capable of implying these costs are significant is a flaw in those models. ACC do not provide much discussion of why they think we should take these costs to be quantitatively important. They also appear to be an artifact of the models' reliance on discrete time, and the timing ambiguities that introduces. As ACC point out, they are using the Svensson, rather than the Lucas and Stokey, timing assumptions. However, these differences in timing assumptions are difficult to evaluate for plausibility, because time, in fact, is not discrete.

The way the timing assumptions operate to create different conclusions is as follows. If M_t/P_t (where the time subscript indicates the date a variable is known and/or chosen) enters a transactions cost term or the utility function, then surprise monetary expansion, with flexible prices, is neutral, though there are still usually real effects from anticipated monetary expansion. Expansion in M is accompanied by an expansion in P that offsets its effects on transactions balances. The simple version of the natural rate hypothesis is thus stood on its head. This result was present in Lucas and Stokey (1983) and has been underlined since, for example, in a paper of mine (1994). If instead it is M_{t-1}/P_t that provides utility or transactions services, then surprise monetary expansions have real effects, but they are purely contractionary in a flexible price model. This is because, with M_{t-1} fixed, any inflationary effects of expansion contract

the availability of transactions services. It is this type of contractionary effect on which ACC rely in generating their results.

But the time unit over which people are stuck with cash, and hence subject to surprise inflation taxes on them, is short on average. It is also endogenous. Thus an increase in the price level can be offset by slightly more frequent trips to the bank, so that the effects on spending of the shrinkage in real balances are slight. Cash-in-advance models, being in their simple form tied to a unit delay between putting aside cash and spending it, cannot account for these possibilities. If one has to choose between M_t/P_t and M_{t-1}/P_t on these grounds, I think the former makes more sense.

In a paper that ACC cite, Nicolini (1998) gives a more extensive discussion of the motivation for this modeling style. He argues that the unit time interval is misleading here; that, in a realistic model with heterogeneous agents that make randomly timed visits to the bank, there will be at any given time some with large balances. He prefers the M_{t-1}/P_t timing (but really, as a long-run modeling strategy, models with heterogeneous agents) because it seems to him a stand-in for missing model realism.

My own view is that Nicolini's argument does not go far enough, because it remains too much focused on cash and transactions technology. The economy is laced with incomplete, nominally denominated contracts, both formal and informal. Surprise inflation has real effects on the terms of these contracts, and a lot of surprise inflation is likely to make people move away from simple nominal contracts. We do, in fact, see the nature of some contracts, for example, labor contracts, shift with the level and variability of inflation. Because nominal contracts reemerge when inflation stabilizes, they apparently are cost effective in some sense. We have little understanding of why these contracts take the form they do, or of the costs of distorting existing contracts by means of inflation, or of inducing people, by means of surprise inflation, to contract differently. These costs may not be very important, but they may also be large. We have no quantitative handle on this issue, and none of the existing approaches to general equilibrium modeling confront the issue.

Nicolini's arguments do not, then, suggest that cash-in-advance or limited-participation models are quantitatively reliable for evaluating the effects of monetary policy. They do, by recognizing the importance of heterogeneity and redistributive effects of inflation, point toward a broader critique of all current approaches to using general equilibrium models for monetary policy evaluation.

Galí's approach justifies its measure of welfare effects by a now-widespread idea, originally given by Rotemberg and Woodford. Under the monopolistic competition assumption, and with the Calvo story about the reason for price stickiness, inflation creates a distribution of prices, with firms that have not for a long time been hit by the exogenous random variable that allows price changes likely to have prices far from the desired level. This leads firms to have dispersed output levels, which is inefficient. But the Calvo story, although a clever modeling idea, has little claim to be grounded in empirical fact. In particular, the costs of inflation it implies rest on firms having no power to

affect the timing of their price changes. Although for some purposes this may be an acceptable simplification, for evaluating welfare this restriction on firm behavior becomes central. If firms can decide to change prices when their prices get far enough out of line, then the distribution of prices, and thereby this source of inefficiency, may be quite insensitive to inflation.

The New Keynesian literature, of course, ends up with a criterion that is a weighted average of variance in inflation and output. Usually, as in Galí's paper, we see tables that show the underlying variance numbers as well as the weighted average that defines welfare. When this practice is followed, we can see how robust are the rankings of policies to changes in the weights. Attempts to translate differences in variances into utility-equivalent differences in mean consumption, though, depend on our taking seriously the derivation of the utility function, and should be treated as speculation.

There is no good argument that the noninflationary level of output would be socially optimal in the absence of random disturbances. One might think that as long as we assume that policy makers can make commitments, so that the possibility of short-run employment gains from surprise inflation does not lead to an inflation bias, it would not make much difference to conclusions if the noninflationary level of output were not optimal; but this is not true. The effects on welfare of variance in inflation and output are second order in the scale of the randomness in the model. In a nonlinear equilibrium model such as those underlying the Galí paper model, randomness affects the mean as well as the variance of the variables in the model. The effects on means are also second order in the scale of the randomness. Unless the effects of changing the means are small, as they are in the neighborhood of the optimal equilibrium, policies cannot be correctly ranked without considering their second-order effects on means. In other words, the log-linearized models that are standard in the literature cannot produce correct rankings of policies.[1]

The defects in the New Keynesian story about welfare costs of inflation and in the ACC story are both cases of a relatively simple, ad hoc rigidity having been suggested as a starting point for modeling nonneutrality, then propagating in the literature because of its convenience as a modeling technique. There is no harm in this process itself, but we need to remain aware that there are many potential ways to generate price stickiness and nonneutrality. Similar qualitative aggregate observations may be accounted for by mechanisms with contradictory implications for welfare evaluation of monetary policy. We therefore need to remain modest in our claims for conclusions about welfare effects of monetary policy. Indeed, it might be best to limit our claims to assertions about the effects of policies on the behavior of aggregate variables. For such claims we have firmer checks against historical data.

[1] This was pointed out in a simple example by Kim and Kim (1999), and there are now a number of individuals or groups preparing and using software that allows second-order expansion of equilibrium models (Collard and Juillard, 2000, Schmitt-Grohé and Uribe, 2001, and Sims, 2000b).

Both sets of authors deal in the papers at hand with policy evaluation in abstract models, at best calibrated roughly to data. In the background is related literature (e.g., Albanesi, Chari, and Christiano, 2000, Clarida, Galí, and Gertler, 2000, Cogley and Sargent, 2001, and Sargent, 1999), some of it by these same authors, which has examined the past half century or so of U.S. monetary policy, evaluating how good it has been, whether it has been improving, and what forces have contributed to its evolution. Although my comments at the conference included some discussion of this literature, particularly of the ACC paper (which is what V. V. Chari actually presented at the conference), here I will simply note that the apparent lack of consensus on the sources of monetary nonneutrality, the dependence of conclusions on assumptions about these sources, and the limited attention to statistical fit in this literature make its conclusions on these issues at best tentative. Despite a widespread belief among economists that the inflation of the 1970s and the return to low inflation in the 1980s and 1990s reflect bad early monetary policy and better later policy, there is little solid empirical support for this view. In fact, there is considerable empirical support for precisely the opposite view: that the inflation and its end were not generated by monetary policy decisions, but rather that monetary policy in fact has had roughly the same systematic component throughout the period.[2]

2. EXISTENCE, UNIQUENESS, AND FISCAL POLICY

The first model in the ACC paper, with flexible prices, arrives at the conclusion that even a monetary authority that cannot make commitments will choose optimal policy, and that that policy is the Friedman rule – contract the stock of non-interest-bearing high-powered money at a rate equal to the real rate of interest. ACC observe that only the real allocation in this equilibrium is unique – there is a continuum of equilibria indexed by the initial price level. However, this observation is given no further discussion.

As in most such models, the non-uniqueness of the initial price level implies the existence of sunspot equilibria, in which the price level is randomly non-unique at every date. There are two real variables that are not uniquely invariant across equilibria: The level of real balances and the level of real lump-sum taxation. Because in reality taxes are not nondistorting, especially if they must be imposed at high rates, it is clear that the policies ACC characterize as "optimal" in this model could not be recommended in practice as good policies. They would require a commitment to tax at unboundedly high rates if realized prices turn out to be very low.

The nonuniqueness ACC find here is essentially the same as that analyzed by Benhabib, Schmitt-Grohé, and Uribe (1998), and it rests on the same somewhat strange type of specification of fiscal policy. Benhabib et al. postulate that

[2] See Hanson (2001), Leeper and Zha (2001), Sims (1999), and Orphanides (2001).

fiscal policy makes the real primary surplus react positively to the real value of total outstanding government liabilities, including high-powered money as well as interest-bearing debt. The ACC fiscal policy, which commits to taxing at a fixed ratio to the real value of the stock of outstanding money (here the only government liability), is a special case. We do not see historical examples of persistent taxation for the sole purpose of contracting the money supply, and current political budget rhetoric does not seem to leave room for this reason to tax. Why should a legislature feel an obligation to increase taxes, with no outstanding debt, when the price level has emerged as lower than expected?

It is more reasonable to suppose that the fiscal authorities make taxes respond, if at all, only to the real value of interest-bearing debt. Furthermore, if we are imagining that the monetary authority is capable of exactly measuring the real rate of return so as to contract M at that rate, we might imagine instead that the fiscal authority (another name for the monetary authority in this model) knows the level of real taxation that is required to contract M at the desired rate when real balances M/P are exactly at the satiation level. If they then formulate policy as a commitment to tax to produce exactly that real primary surplus, regardless of what happens to prices, and the monetary authority commits to maintaining a zero nominal interest rate, then the ACC model will have a Friedman-rule equilibrium with a uniquely determined price level.

The Friedman rule is still a questionable policy, though. If the tax authority sets the primary surplus too low, the policy is unsustainable, and whereas the uncertainty about how the policy will become sustainable persists, high inflation is likely to result. The surplus could be set high enough to allow a margin of error – high enough to contract a volume of real balances well exceeding the level that satiates transactions demand at a rate above any value the real discount factor is likely to take on. However, if we recognize a cost to high taxation, this policy choice is unattractive.

If these transactions costs considerations are the primary reason for worrying about inflation, then the best policy in practice is likely to be setting a low positive nominal rate. This, by fixing velocity uniquely, will, in combination with any primary surplus above some minimum level, guarantee a unique equilibrium. Such policies do require fiscal support. A steady flow of tax revenue is devoted to contracting the money stock. However, widespread existing conventions, in which central banks "own" some of the government's debt and can use the flow of interest income for policy purposes without legislative second-guessing, would support the necessary flow of fiscal resources.[3]

[3] Actually, the situation in the United States, where the central bank backs its high-powered money mainly with holdings of nominal domestic government debt, is not common. Where the central bank balance sheet includes a large component of real assets or of foreign currency denominated securities, the current conventions of central bank independence would not suffice to provide fiscal support for such a mildly deflationary policy. Explicit fiscal backing would be required (Sims, 2000a).

Specifying monetary policy not as a fixed nominal rate but instead as contraction of the nominal money stock at a fixed rate $g > \beta$ in this model is not compatible with commitment to a fixed real primary surplus. In the multiple equilibria that ACC find for this case, in all but one, real balances grow at the rate g/β. This requires, with their specification of fiscal policy, that real primary surpluses grow without bound. With the primary surplus instead fixed in real terms, the shortfall has to be made up by borrowing, and the borrowing would have to produce real interest-bearing debt growth at the rate β^{-1}, which is incompatible with private transversality. The equilibrium in which prices and M both grow at rate g, so M/P remains constant, determines a unique nominal rate, thus a unique initial M/P, and thus a unique initial P. However, in general, this initial P will not be the value that matches initial $(B + M)/P$ to the discounted present value of primary surpluses, so there is no corresponding equilibrium. In other models, where transactions demand is highly interest sensitive and a barter equilibrium exists, the fixed M, fixed surplus combination is actually consistent with equilibrium – an indeterminate continuum of them, all of which involve inflationary dissipation of real balances and convergence to barter equilibrium.

The uniqueness issues I raise here in connection with the first of the three ACC models in the paper in this volume are more complicated to analyze in the sticky price and limited-participation models, and ACC do not fully unravel them. They consider only recursive equilibria with certain natural candidate state vectors, which leaves open the question of whether there may be equilibria indexed by, for example, lagged prices. Uniqueness issues are brought out, but in the same sense also incompletely analyzed, in the paper by the same authors presented at the conference (Albanesi et al., 2000). There they suggest that the non-uniqueness that can arise in a model that allows more elastic response of transactions technology to interest rates might explain postwar U.S. monetary history. In that paper, it would be even more important to recognize that jumps between equilibria may imply substantial jumps in the primary surplus with their assumptions about fiscal policy, and also the possibility that commitment to a real primary surplus can eliminate indeterminacy.

The Galí paper also ignores the fiscal aspect of monetary policy. This leads to its assertion that a fixed nominal interest rate implies indeterminacy of the price level if interpreted literally. This is not true if the fixed nominal interest rate is accompanied by a commitment to a real primary surplus. It would be simpler to describe the fixed interest rate policy this way. As the paper now describes it, it sounds difficult to implement, whereas in fact it would be easy.

In a model that recognized the revenue from seignorage, the fixed interest rate policy implemented as Galí describes it here (via a restricted variant on a Taylor rule, with fiscal policy implicitly making the primary surplus responsive to the real debt) would probably not deliver the same equilibrium as a nominal interest rate peg accompanied by a primary surplus peg. However, in the model as laid out here, the two versions of a fixed nominal rate policy probably do deliver equivalent equilibria. My own experiments suggest that, if a disturbance

is introduced into the Phillips curve as Galí does later in the paper, an interest rate peg can deliver lower inflation variance (though still higher output variance) than the Taylor rule in this model. It would be a good idea, before taking the policy evaluations displayed here as the last word even for this type of model, to consider the effects of such shocks on the rankings of policy rules.

3. THE NEW PHILLIPS CURVE

Simple New Keynesian models use a purely forward-looking "Phillips Curve" derived from Calvo-style price adjustment. These models imply that inflation "leads" the output gap, and Fuhrer and Moore (1995), in a careful two-equation analysis, showed that in what was then their usual form these models are qualitatively at variance with the time-series facts. In another paper, working with a formally similar inventory model, Fuhrer, Moore, and Schuh (1995) show that, for these models, likelihood-based multiple-equation methods are much more reliable than inference based on instrumental variables.

The literature Galí cites on the newest formulation of the New Phillips Curve (Galí and Gertler, 1999; Sbordone, 1998) retreats from the standard of care in empirical evaluation set by the Fuhrer–Moore paper, using instrumental variables methods alone for estimation and using informal measures of fit that can easily be misleading.

To understand the problems associated with the approach in these papers to assessing fit, consider the New Phillips Curve model laid out in the Galí–Gertler (GG) paper. It specifies the same Phillips Curve that appears in this volume's Galí paper:

$$\pi_t = \beta E_t \pi_{t+1} + \kappa x_t, \tag{3.1}$$

with x_t interpreted as being the log of the labor income share in the nonfarm business sector. This is, on the face of it, an equation that can be estimated by an ordinary least squares (OLS) regression of π_{t+1} on π_t and x_t. Without explaining why, GG estimate it instead by instrumental variables, normalized as in (3.1), and they omit π_t from the instrument set.[4]

Solving this equation forward in isolation, they obtain an implied relation between π_t and a discounted sum of expected future values of x:

$$\pi_t = \kappa \sum_{t=0}^{\infty} \beta^s E_t x_{t+s}. \tag{3.2}$$

Then they apparently[5] use an unrestricted bivariate vector autoregression (VAR) to form expected future values of x, substitute them into the right-hand side

[4] The normalization would not have mattered if they had kept π_t in the instrument set, as in that case instrumental variables estimates would have been identical to those from OLS.

[5] The nature of the forecasting rule used to form $E_t x_{t+s}$ is not completely clear from the paper. All that is stated explicitly is that it is based on data on π and x dated t and earlier.

of (3.2), and call the result "fundamental inflation." The theory predicts that fundamental inflation defined this way is exactly observed inflation. This is the intuition behind plotting the two, noting how closely they track each other, and counting what appears to the eye a close tracking as support for the theory.

However, here we are not testing a theory in which an explanatory variable x is claimed to be important, whereas under an alternative view x is expected to be unrelated to π. When the gap is measured as the labor share, it is being measured as a variable that any reasonable macro theory would treat as highly endogenous, affected by all kinds of structural disturbances. It is not surprising that linear combinations of it with π should turn out to be highly correlated with π. The theory claims that fundamental inflation should be identical to actual inflation, and this is a simple linear restriction on the coefficients of the VAR. The theory should be tested directly as such a restriction. To understand this point, it may help to write out the unrestricted VAR as

$$\begin{bmatrix} \vec{\pi}_t \\ \vec{x}_t \end{bmatrix} = z_t = Az_{t-1} + \varepsilon_t, \tag{3.3}$$

where $\vec{\pi}_t$ and $\vec{x}_t$ are $k \times 1$ vectors containing current and lagged values of π and x. In this form, the system allows us to write

$$E_t z_{t+s} = A^s z_t, \tag{3.4}$$

and therefore, according to theory,

$$\pi_t = e_1 z_t = e_{k+1} \kappa E_t \left[\sum_{s=0}^{\infty} \beta^s A^s z_t \right] = (I - \beta A)^{-1} z_t, \tag{3.5}$$

where e_i denotes the unit vector with a one in its ith position. This implies

$$e_1 = e_{e+1} \kappa (I - \beta A)^{-1} \tag{3.6}$$

or, equivalently,

$$e_1 \beta A = e_1 - \kappa e_{k+1}. \tag{3.7}$$

Note that this is simply the assertion that the first equation of the VAR should be exactly (3.1), normalized to have π_{t+1} on the left with a unit coefficient, and with the E_t operator dropped. The theory should be evaluated by examining, using likelihood-based measures, the validity of this straightforward linear restriction on A.[6]

[6] Comparing the time-series behavior of the left- and right-hand sides of (3.2) is widespread in applications of what is known as the Campbell–Shiller methodology. Such time-series plots can be useful in assessing a model, but they should not be treated as measures of fit. Many authors construct formal tests of (3.6) as a nonlinear restriction on A. As has recently been pointed out by Mercereau (2001), there is good reason to think that inference about $(I - A)^{-1}$ in these models, based on asymptotics, will be much less reliable than inference about A itself.

Even if all the inference were done more carefully, however, we should bear in mind that we are not directly interested in restrictions within a model incorporating a single-gap measure. We are interested in comparing models. We are interested in how a full-fledged model works in providing forecasts and policy analysis when incorporating one or the other measure of the "gap," forward- or backward-looking specifications, and so on. As yet, the literature arguing for a Phillips Curve with labor share standing in for the gap has given us no evidence on these questions, even on the simple one of whether forecasts of a multivariate model are improved by a theoretically constrained addition of a labor share variable.

4. CONCLUSION

In emphasizing how tenuous and loosely linked to real data are the monetary policy evaluations in these papers, I do not intend to imply that this is uninteresting work. Precisely because we still understand so little in this area, work with stylized, unrealistic models can be interesting and worthwhile. Yet, in central banks around the world, there is increasing recognition of the value of being more explicit and quantitative about planned time paths of target variables such as inflation, output, and employment and about how current policy actions are expected to influence these time paths. Accomplishing this does not require models that explain why people hold money and why money is nonneutral, but it does require multivariate models grounded in data as well as in theory. Although these two papers did not set out to provide or improve such models, they are representative of a great deal of academic macroeconomic research. It seems to me that work on improving models that contribute to current problems in implementing monetary policy deserves relatively more attention from academic macroeconomists.

References

Albanesi, S., V. V. Chari, and L. J. Christiano (2000), "Expectation Traps and Monetary Policy: Preliminary Draft," Discussion Paper, Northwestern University, available at www.faculty.econ.northwestern.edu/faculty/christiano/.

Benhabib, J., S. Schmitt-Grohé, and M. Uribe (1998), "Monetary Policy and Multiple Equilibria," Discussion Paper, New York University.

Clarida, R., J. Galí, and M. Gertler (2000), "Monetary Policy Rules and Macroeconomic Stability: Evidence and Some Theory," *Quarterly Journal of Economics*, 115, 147–180.

Cogley, T. and T. J. Sargent (2001), "Evolving U.S. Post-World War II Inflation Dynamics," *NBER Macroeconomics Annual.*

Collard, F. and M. Juillard (2000), "Perturbation Methods for Rational Expectations Models," Discussion Paper, CEPREMAP, Paris, available at fabrice.collard@cepremap.cnrs.fr.

Fuhrer, J. C. and G. R. Moore (1995), "Inflation Persistence," *Quarterly Journal of Economics*, 110, 127–159.

Fuhrer, J. C., G. R. Moore, and S. D. Schuh (1995), "Estimating the Linear-Quadratic Inventory Model: Maximum Likelihood Versus Generalized Methods of Moments," *Journal of Monetary Economics*, 35(1), 115–157.

Galí, J. and M. Gertler (1999), "Inflation Dynamics: A Structural Econometric Analysis," *Journal of Monetary Economics*, 44, 195–222.

Hanson, M. (2001), "Varying Monetary Policy Regimes: A Vector Autoregressive Investigation," Discussion Paper, Wesleyan University.

Kim, J. and S. Kim (1999), "Spurious Welfare Reversals in International Business Cycle Models," Discussion Paper, Brandeis University, available at http://www.people.virginia.edu/∼jk9n/.

Leeper, E. and T. Zha (2001), "Modest Policy Interventions," Discussion Paper, Indiana University and Federal Reserve Bank of Atlanta, available at http://php.indiana.edu/∼eleeper/Papers/lz0101Rev.pdf.

Lucas, R. E. Jr. and N. Stokey (1983), "Optimal Fiscal and Monetary Policy in an Economy without Capital," *Journal of Monetary Economics*, 12(1), 55–93.

Mercereau, B. (2001), "Does Wall Street Matter? Impact of Stock Markets on the Current Account Dynamics," Discussion Paper, Yale University.

Nicolini, J. P. (1998): "More on the Time Consistency of Monetary Policy," *Journal of Monetary Economics*, 41(2), 333–350.

Orphanides, A. (2001), "Monetary Policy Rules, Macroeconomic Stability, and Inflation: A View from the Trenches," Discussion Paper, Board of Governors of the Federal Reserve System.

Sargent, T. J. (1999), *The Conquest of American Inflation*. Princeton, NJ: Princeton University Press.

Sbordone, A. M. (1998), "Prices and Unit Labor Costs: A New Test of Price Stickiness," Discussion Paper, Rutgers University, available at http://fas-econ.rutgers.edu/sbordone/.

Schmitt-Grohé, S. and M. Uribe (2001), "Optimal Fiscal and Monetary Policy Under Sticky Prices," Discussion Paper, Rutgers University and University of Pennsylvania.

Sims, C. A. (1994), "A Simple Model for Study of the Determination of the Price Level and the Interaction of Monetary and Fiscal Policy," *Economic Theory*, 4, 381–399.

Sims, C. A. (1999), "Drift and Breaks in Monetary Policy," Discussion Paper, Princeton University; presented at a plenary session of the July, 1999 meetings of the Econometric Society, Australasian region; available at http://www.princeton.edu/∼sims/

Sims, C. A. (2000a), "Fiscal Aspects of Central Bank Independence," Discussion Paper, Princeton University, available at www.princeton.edu/∼sims.

Sims, C. A. (2000b), "Second Order Accurate Solution of Discrete Time Dynamic Equilibrium Models," Discussion Paper, Princeton University, available at eco-072399b.princeton.edu/gensys2/.

CHAPTER 6

Consumption Smoothing and Extended Families

Orazio P. Attanasio and José-Víctor Ríos-Rull

1. INTRODUCTION

Agricultural economies are characterized by substantial fluctuations in individual income. Some of these fluctuations affect all members of the society, whereas others are individual specific. Income fluctuations do not have to translate into consumption fluctuations (that people abhor given convex preferences). However, all too often, income fluctuations induce (perhaps mitigated) consumption fluctuations. Moreover, there is evidence that idiosyncratic risk is not fully insured even within relatively small and closed groups. Udry (1994), for instance, discussing his evidence from rural Nigeria, states that "it is possible to reject the hypothesis that a fully Pareto-efficient risk-pooling allocation of village resources is achieved through these loans. The mutual insurance network available through these loans to households in rural northern Nigeria is important, but it is incomplete" (p. 523).

We think of agricultural societies (villages and islands) as both having undeveloped financial markets and being small enough so that anonymity does not exist within the village. At the same time, however, information about idiosyncratic shocks could be difficult to convey to the outside world. In other words, these societies might have limited enforcement capability. The lack of developed financial markets prevents the members of these societies from borrowing and lending and from insuring both among themselves and with the outside world. The fact that, within the economy (or smaller subsets of agents), information problems are negligible, while enforceability problems might be serious, suggests the modeling framework to use when thinking of what type of institutions may develop to substitute for the missing markets. In short, the only type of arrangement that may be possible is self-enforcing contracts, that is, contracts that are sustained by the mutual interest of the parties of maintaining the relationship.

Here we discuss three related issues that pertain to the extent to which is what we define as island or village economies, which are just small, isolated, poor agricultural societies, can smooth consumption. We start from the empirical implications of perfect insurance and discuss the empirical evidence that shows

that consumption does fluctuate substantially, and more than what is implied by perfect insurance. However, the empirical work we describe also shows that, despite large consumption fluctuations, income fluctuations are even larger, indicating the existence of some smoothing mechanism. The evidence points to nonformal channels through which this insurance is carried out. To do the empirical assessment of consumption assessment in village economies, we offer a brief overview of the implications of first best.

Our second theme is the analysis of computable models that show the extent to which extended families can achieve insurance, sometimes total insurance, but usually partial insurance. We focus on self-enforcing contracts and document the implications of various characteristics of the environment in shaping the amount of possible insurance that can be implemented within the extended family. In particular we consider the effects of both preferences and income processes for the amount of risk sharing that can be sustained in equilibrium. We also discuss the empirical implications of this class of models and the scant empirical evidence on them.

Our third theme is normative in nature. We are interested in understanding the extent to which policy can help in increasing the welfare of the members of village societies. Specifically, we have in mind institutions (henceforth called World Banks) that can partially observe the realizations of the shocks that affect income (what we call the aggregate part, the part that is common to all villagers) and that can compulsorily provide insurance against these fluctuations. This, in principle, sounds like a good idea if only because the villagers do not have access to this type of insurance in the open market. What we find particularly interesting is that this type of policy has a deep influence over the type of enforceable arrangements, the social fabric, that the villagers can make for themselves. First, we show examples in which the well-intentioned policy of the World Bank may induce a reduction of welfare. The reason for this apparent surprise could be thought of as a special case of Hart's standard result, in which opening markets but staying shy of complete markets may reduce welfare. We think of an interpretation in terms of the destruction of the social fabric that occurs when the public policy starts: with certain forms of government insurance, the incentives for a private insurance scheme to battle the consequences of autarky are smaller. They are so much smaller that, in certain circumstances, they may completely compensate for the direct good effect of the policy.

We then turn to what we think is the key issue, which is how to design policies that do not have the aforementioned problem. Moreover, can we construct policies with the property that they strengthen the social fabric of self-enforcing contracts rather than cripple it? Even more, can we do it in a simple way that can actually be carried out by a middle-of-the-road public institution? Whereas we think that the first objectives are relatively self-evident, the latter one requires some comments. The reader has probably noted the similarity of the language with that of the literature in implementation theory. However, although we use some ideas and tools of implemention theory, our focus is quite different. The implementation literature provides a number of theoretical

results that are often quite abstract. Moreover, specific outcomes are often implementable only through very complex and sometimes esoteric mechanisms, which we see as unfit for de facto implementation by a bureaucratic agency. We place the emphasis not on proving the implementability of the first best with some mechanism, but on achieving relatively good outcomes with very simple mechanisms.

In studying the properties of simple mechanisms, we use our and, very often, our colleagues' intuition for what could work. However, the ultimate judge of the goodness of a policy or mechanism is the actual set of equilibria associated with it. To find this set, we use computational methods. Computing the equilibria, rather than characterizing them, allows us to say many things about a small class of economies rather than a few things about a larger set of economies. This means that, ultimately, to know how good the performance is of a specific policy, we have to go to the details of the environment where the policy is put in place.

One problem we face in the construction of simple mechanisms is that, often, the equilibrium we would like is not necessarily unique. This problem arises from the fact that the government can only manipulate aggregate payments and that the "punisher" does not necessarily have a strong incentive to deprive the other agent of the aggregate payment, as this would not affect his or her utility. To circumvent this problem, we try to construct mechanisms in which, off the equilibrium, an agent has strong incentives to implement the punishment.

2. THE PERFECT INSURANCE CASE: THEORETICAL AND EMPIRICAL IMPLICATIONS

In what follows, we consider economies in which individuals can enter contracts to diversify idiosyncratic risk and therefore smooth consumption. Individuals within these economies belong to what we call extended families, whose existence and membership are exogenously given. We start our discussion by illustrating the theoretical and empirical implications of a model where first best can be achieved. By first best we mean, however, the first-best allocation within the members of the extended family. We do not consider the possibility that members of the extended family share risk with people outside it and/or that new extended families are formed. A possible justification of this assumption is that the absence of information among members of different families prevents intertemporal trade among them completely. Obviously, this is a simplification. Although we work with extended families made of two individuals, the size of the family is not particularly important. Therefore, if one thinks of the family as being as large as the village, one can get the standard first-best results analyzed in the literature.

The model we present in this section is useful for providing a benchmark and for introducing notation. We consider simple cases, namely endowment economies, possibly without storage possibilities. Although many of the results can be generalized to more complex situations, we find it useful, for explanatory

reasons, to discuss the main ideas by using a simple model. Along the way, we discuss which extensions to more realistic and complex settings are likely to affect our results.[1]

In addition to the introduction of the model, we also discuss its empirical implications and the main empirical findings available in the literature. As we document substantial deviations from first-best allocations in village economies, this part constitutes a motivation for considering models where the first best is not achieved for very specific reasons.

2.1. The Basic Model

Consider an exchange economy populated by many individuals. These individuals receive endowments that are functions of idiosyncratic and aggregate shocks. Within the economy there exist extended families, exogenously determined and of fixed size. What identifies the families is the fact that members of the family have perfect information about each other's idiosyncratic shocks. However, we should note that the size of what we call the extended family can easily be enlarged. For simplicity, we assume that extended families are made of two individuals.

Let z denote the aggregate shock with finite support in Z. This shock is common to all individuals in the economy. Furthermore, the shock z is Markov with transition matrix $\Gamma_{z,z'} = \text{prob}(z_{t+1} = z' | z_t = z)$, and stationary distribution γ_z^*.[2] Let $s \in S$ denote the idiosyncratic or individual shock, which is also Markov, and which is specific to each household. Note that s may be multivalued, so that it can incorporate both temporary and permanent elements and also has finite support. Conditional on two consecutive realizations of the aggregate shock,[3] we write the stochastic process for s as having transition $\Gamma_{s,z,z',s'} = \text{prob}(s_{t+1} = s' | z_{t+1} = z', z_t = z, s_t = s)$, and unconditional means $\bar{z}$ and $\bar{s}$. In each state $\{z, s\}$ agents get endowment $e(z, s)$. We compactly write $\epsilon \equiv \{z, s\}$ and its transition $\Gamma_{\epsilon,\epsilon'}$. We use the compact notation $y = (z, s_1, s_2)$, and we refer to its components as $\{z(y), s_1(y), s_2(y)\}$, which are the aggregate shock and the idiosyncratic shock of agents 1 and 2, respectively. We also compactly write the transition matrix of the pair as $\Gamma_{y,y'}$ We denote by $\gamma^*(y)$ the stationary distribution of the shocks.[4] Moreover, the history of shocks up to t is denoted by $y^t = \{y_0, y_1, \ldots, y_t\}$. We use $\pi(y^t | y_{-1})$ to denote the probability of history y^t conditional on the initial state of the economy y_{-1}.

In the absence of enforceability (and information) problems, the members of the extended family can share risk and achieve a welfare improvement, if

[1] The model we use is based on that from Attanasio and Ríos-Rull (2000a).

[2] There are simple conditions that we assume and that guarantee that the stationary distribution exists, is unique, and is the limit for any initial condition.

[3] See Castañeda, Díaz-Giménez, and Ríos-Rull (1998) for details about the modelization of joint aggregate and idiosyncratic shocks.

[4] We make sufficient assumptions on the Γ's to ensure that there is a unique stationary distribution and no cyclically moving subsets.

utility functions are concave. The characterization of the allocation of resources under this sort of arrangement can be described by looking at a central planner problem. This was done by Townsend (1994). In particular, with two members in the extended family, the planner's problem maximizes a weighted sum of utilities subject to resource constraints. That is, for nonnegative weight λ_1, with $\lambda_2 = 1 - \lambda_1$, the planner chooses an allocation $\{c_1(y^t), c_2(y^t)\}$ for all y^t to solve

$$\max_{\{c_i(y^t)\}} \lambda_1 \sum_{t=0}^{\infty} \sum_{y^t} \beta^t \pi(y^t) u[c_1(y^t)] + \lambda_2 \sum_{t=0}^{\infty} \sum_{y^t} \beta^t \pi(y^t) u[c_2(y^t)] \tag{2.1}$$

subject to the resource constraints

$$c_1(y^t) + c_2(y^t) = e_1(y^t) + e_2(y^t). \tag{2.2}$$

The solution to this problem has to satisfy the following condition:

$$u'[c_i(y^t)]\lambda_i = \xi(y^t), \tag{2.3}$$

where $\xi(y^t)$ is the Lagrange multiplier of the resource constraint after history y^t. This condition can be rewritten as

$$\frac{u'[c_1(y)]}{u'[c_2(y)]} = \frac{\lambda_2}{\lambda_1}. \tag{2.4}$$

Equation (2.4) poses a very strong restriction on the equilibrium allocations: the ratio of marginal utilities is constant across all periods and states of nature, which allows us to drop the whole history as an argument of the consumption choices: only the current state y affects the consumption allocation. This is true for all possible weights that the planner may be using. This property is the one that we will like to find in the data as evidence of the agents being able to insure against shocks.

There is another important property in the characterization of the allocations provided by the planner problem: that theory by itself does not predict which specific allocation will occur. In large economies, where competitive equilibrium is a good representation of agents' interactions, there are strong additional restrictions on which set of weights are associated with the allocations that are picked. Essentially, there is generically at most a finite number of allocations that can be implemented as competitive equilibria without transfers. However, in families, or more generally in economies with a small number of agents, we do not know how the agents will share the gains from trade.

2.2. Empirical Implications and Evidence

Equation (2.3) was stressed in a seminal paper by Townsend (1994), after which the implications of full risk sharing were studied by several authors. Townsend (1994) proposed testing the hypothesis that changes in individual (log) consumption, after controlling for changes in aggregate consumption, are not correlated with changes in the level of resources available to an individual. The idea

is that (log) consumption can be taken to approximate the log of marginal utility, whose change should reflect only changes in the resource constraint multiplier in the central planner problem. This can be seen by taking time differences of the log of Equation (2.3): the differencing eliminates the unobserved Pareto weight.

Townsend (1994) tested this important implication of first-best allocation by using the International Crops Research Institute for the Semi-Arid Tropics (ICRISAT) data from semiarid India, with mixed empirical results. Mace (1991) and Cochrane (1991) implemented similar tests on U.S. data. In particular, Mace (1991) used Consumer Expenditure Survey data, whereas Cochrane (1991) used Panel Study of Income Dynamics (PSID) data. Hayashi, Altonji, and Kotlikoff (1996) also used PSID data to test insurance both across and within families. Overall, the evidence suggests considerable rejections of the perfect insurance of aggregate shocks. Attanasio and Davis (1996) use grouped (by year of birth cohort and education) consumption data from the Consumption Expenditure Survey (CEX) together with wage data from the Current Population Survey (CPS) and find strong rejections of the null, especially when low-frequency changes are considered.

An alternative test of the null was recently proposed by Attanasio, Blundell, and Preston (2000), who looked at the variance of log consumption within certain groups. This test is based on the idea that the variances of the marginal utility of consumption should be constant over time in a group that insures idiosyncratic shocks of its members, as it should reflect only the variance of the Pareto weights. This can be seen clearly from Equation (2.4).

There are two advantages to this test. First, by considering the within-group variance, one can test insurance *within* a group, whereas the group means used by Attanasio and Davis (1996) tested for insurance of shocks across groups. Second, by changing the definition of groups, one can focus on different types of shocks that can be economically meaningful. Attanasio et al. (2000) find much stronger rejections of the null for broadly defined groups, such as those defined only by year of birth cohorts, than for groups defined by cohort and education.

The evidence both in Attanasio and Davis (1996) and in Attanasio et al. (2000) is consistent with the hypothesis that transitory, high-frequency shocks are somehow smoothed out, whereas more permanent relative shocks are reflected in changes in relative consumption. The idea that permanent components are more difficult to smooth out is consistent both with the idea that consumers can save only to smooth out shocks, and with the fact that permanent shocks are more difficult to share in the presence of imperfect enforceability.

An important point to note is that of the power of the tests of perfect insurance. This issue is particularly relevant when the estimates of the coefficients used to test the null are likely to be affected by attenuation bias because of measurement error. Some of the results Mace (1991) obtained, which indicated a nonrejection of the null, might be a consequence of measurement error in income. Attanasio and Davis (1996) and Attanasio et al. (2000) try to get around this problem by grouping and instrumenting. However, the difficulty in constructing a suitable instrument might explain some of the high-frequency

results that Attanasio and Davis (1996) get. Indeed, distinguishing between transistory shocks that can easily be self-insured and measurement error can be quite hard.[5]

With the exception of Townsend (1994), the papers cited herein use data from developed countries. However, many other studies after that of Townsend have looked at the implications of perfect insurance in developing countries. These include the studies by Ravallion and Chaudhuri (1997) and Atkeson and Ogaki (1996). Atkeson and Ogaki in particular, consider a Stone–Geary utility function. More recently, Ogaki and Zhang (2001) failed to reject the null of perfect insurance, once they allowed for a surviving level of consumption.

It is interesting to note that most of these tests look only at consumption (and income) realizations, without necessarily requiring specific information on the instruments that people might use to smooth out idiosyncratic fluctuations. It has been suggested that, in village economies characterized by limited storage capabilities, small information problems and repeated interactions among agents, insurance contracts, and arrangements are extremely rare.[6] Instead, people enter what Platteau (1997) has defined as quasi-credit arrangements, which are by their nature quite similar to the type of contracts, halfway between credit and insurance, that we just described. Platteau and Abraham (1987) and Platteau (1997) find extensive evidence in this respect from fisherman villages in Southern India and Africa. To study risk sharing, therefore, it can be quite profitable to study, in addition to income and consumption allocation, the particular instruments that households use to (partly) insure income shocks, such as gifts, informal credit, transfers, and so on, when data on these variables exist.[7]

In an important paper, Udry (1994) considers credit arrangements in Nigeria and shows that these arrangements are not consistent with the implications of full risk sharing. In particular, Udry finds that both the maturity and the effective interest rate on loans (which determine the terms of repayment) often depend on the shocks that affect the two parties of the contract. Not only does he find that borrowers affected by negative shocks pay back less, but also that lenders in the same situation get paid back more! Udry also discusses and estimates models in which information flows between the partners are not perfect (but shocks are observed ex post by a village authority) and some households rationally default on their loans in some state of the world, even though they are punished in this

[5] A possibility is to use different data sources that report measures of the same variable. Attanasio et al. (2000) use the CPS wage to instrument CEX wages. If the measurement error is the only problem, this procedure should provide a reliable solution to it.

[6] Besley (1995) provides an interesting survey of risk-sharing institutions and credit arrangements in developing countries. Fafchamps (1999) provides an interesting survey of the available evidence on quasicredit.

[7] In a recent paper, Dercon and Krishnan (2000) test first-best allocation of resources within households. Using data from rural Ethiopia they marginally reject efficient risk sharing among husbands and wives. Interestingly they also estimate the Pareto weights implied by their data and relate them to observable characteristics that might proxy for the bargaining power in the marriage.

case by the village authority.[8] Udry estimates such a model by using data on debts, repayments, and defaults by maximum likelihood. The fact that there he has only one cross section implies that he is forced to consider a two-period model and cannot explicitly consider the dynamic effects induced by imperfect enforceability in a repeated context. However, estimating a model of bilateral loan contracting that also considers the possibility of default, he finds that "borrowers and lenders are engaged in risk pooling through state-contingent loan repayments." Such schemes bear remarkable similarities to the model we discuss.

Fafchamps and Lund (2000) discusses the importance of informal transfers and gifts in risk-sharing agreements in rural Philippines. Like Udry (1994), he sets his empirical analysis as a test of the perfect insurance hypothesis and explicitly mentions imperfect enforceability and imperfect information as a possible explanation of the rejections of the null. Interestingly, Fafchamps has information on "network" membership and can test efficient risk sharing not just at the village level but at the network level as well.

Interestingly, both Udry (1994) and Fafchamps and Lund (2000) do not use information on consumption, like in many of the perfect insurance tests mentioned herein. Instead, in addition to the information on income shocks, they use data on the instruments used for consumption smoothing: (informal) credit in the case of Udry and transfers and gifts in the case of Fafchamps.

3. THE FAILURE OF PERFECT INSURANCE

Because the implications of perfect insurance are often empirically rejected, we move on to the discussion of models in which the failure of perfect insurance is modeled explicitly. We start our discussion with a brief mention of models with imperfect information. However, the focus of this section and of our paper is on models in which perfect insurance fails because of imperfect enforceability of contracts. We believe that these models are particularly suitable for analyzing situations in which there are limited storage capabilities (therefore reducing the possibility of self-insurance) and in which information about idiosyncratic shocks is reasonably public within the village. In contrast, it might be difficult to convey this type of information to the external world and therefore enforce punishments for deviations from preestablished contracts. This information structure is important for the construction of optimal aggregate insurance schemes, which we discuss in Section 6.

The extent to which the models we discuss are relevant from a policy perspective is largely an empirical question. It is therefore important to focus on the implications of the models we consider and to discuss the available empirical evidence. For this reason, we conclude the section with a discussion of the empirical evidence on models with imperfect enforceability.

[8] Moreover, Udry (1994) assumes that there are some transaction costs in loans, so that there is a positive mass at zero loans.

3.1. Imperfect Information

There is a large body of work that attempts to understand what can be done when first best is not implementable. This group includes research by Atkeson and Lucas (1992), among many others, that studies the implementable allocations when the current shock of agents (which can be either an income or a preference shock) is unobservable.[9] In many cases, it turns out that the optimal allocation is characterized by ever-increasing inequality.

In a very recent and nice paper, Cole and Kocherlakota (2001) have shown that, when storage is both feasible for the agents and unobservable by third parties, the optimal implementable allocations are essentially those that can be achieved by agents' holding an asset, perhaps in negative quantities, that has a rate of return that is not state contingent. In this sense Cole and Kocherlakota have provided an important link between the literature that studies constrained optimal allocations and other literature that is interested in the properties of allocations and prices when the market structure is incomplete. Models with only one asset and with numerous agents differing in income and wealth have been used to address various questions in macroeconomics (Aiyagari, 1994, Castañeda et al., 1998, 2002, Huggett, 1993, and Krusell and Smith, 1998). Models of this type and with more than one asset yet still incomplete markets have been used in finance to look for solutions to the equity premium puzzle (Krusell and Smith, 1997, with a large number of agents, and Heaton and Lucas, 1992, Marcet and Singleton, 1990, and Telmer, 1992, in economies with few agents).

We think that obervability issues are inherently related to the anonymity of large societies, and that for small agricultural economies it is more useful to organize our thinking around the problem of enforceability. This is especially true if we are interested in economies where the members have difficulties storing assets. Accordingly, in the rest of the paper we consider economies where agents cannot store goods or hold assets and where there are enforceability but not observability problems.

3.2. Imperfect Enforceability

Here we consider environments where there is no access to a technology to enforce contracts, which may preclude agents from achieving first-best allocations. We want to think of this as a situation where agents cannot convey information about shocks in an easily verifiable way to the outside world. In such a situation, agents are likely to enter only these contracts that are self-enforceable. In particular, look at allocations that can be achieved within a repeated game where there is a threat to revert to the worst-possible subgame perfect equilibrium. This is what we call the "autarky" equilibrium, that is, one

[9] Other relevant papers include those by Phelan and Townsend (1991), Green (1987), and Wang and Williamson (1996).

in which each member of the extended family consumes her or his idiosyncratic endowment. This approach arises from the work of Abreu (1988) and Abreu, Pearce, and Stacchetti (1990) and has been used by most papers in the literature. The list of papers that have looked at problems such as these (which are essentially consumption smoothing problems) include those by Thomas and Worrall (1990), Ligon, Thomas, and Worrall (2000, 2001), Alvarez and Jermann (1998), and Kocherlakota (1996).[10] Kehoe and Levine (1993) have looked at this problem as it relates to access to markets.

We look at allocations that are accepted voluntarily by agents that in every period and state of nature have the option of reverting to autarky.[11] Therefore, if we denote by $\Omega(\epsilon)$ the value of autarky, our assumptions on the endowments processes imply that it can be written as

$$\Omega(\epsilon) = u[e(\epsilon)] + \sum_{\epsilon'} \Gamma_{\epsilon,\epsilon'} \Omega(\epsilon'). \tag{3.1}$$

Given this, we find that the enforceability constraints we have to consider are given by the following expression:

$$\sum_{r=t}^{\infty} \sum_{y^r} \beta^{r-t} \pi(y^r | y^t) u[c_i(y^r)] \geq \Omega[\epsilon(y_t)]. \tag{3.2}$$

3.2.1. A Recursive Formulation

These constraints (one for each of the two agents) can be used to change the central planner problem. Adding these two constraints to the maximization problem just given changes the nature of the problem considerably. In particular, the problem becomes, as it is written, nonrecursive. We follow Marcet and Marimon (1992, 1995) and rewrite the planner problem to make it recursive (see also Kehoe and Perri, 1997). In particular, we can write the Lagrangian for such a problem as

$$\sum_{t=0}^{\infty} \sum_{y^t} \beta^t \pi(y^t) \left\{ \sum_{i=1}^{2} \lambda_i u[c_i(y^t)] + \sum_{i} \mu_i(y^t) \left[\sum_{r=t}^{\infty} \sum_{y^r} \beta^{r-t} \pi(y^r | y^t) u[c_i(y^r)] - \Omega_i[\epsilon(y_t)] \right] \right\}, \tag{3.3}$$

[10] Coate and Ravallion (1993) were among the first to consider self-enforcing contracts. However, they restrict themselves to static contracts that are not necessarily optimal.

[11] As Ligon et al. (2000) do, we could also subtract from the value of autarky any punishment that can conceivably be imposed on an individual that deviates from the preagreed contract. As we do not use these punishments here, we do not consider them in the equation here to avoid clustering the notation.

plus the standard terms that relate to the resource constraints. The μ_i are the multipliers associated with the participation constraints. Noting that $\pi(y^r|y^t)$ can be rewritten as $\pi(y^r) = \pi(y^r|y^t)\pi(y^t)$, we can rewrite the Lagrangian as

$$\sum_{t=0}^{\infty}\sum_{y^t}\sum_{i}\beta^t\pi(y^t) \times \left\{M_i(y^{t-1})u[c_i(y^t)] + \mu_i(y^t)\left[u[c_i(y^t)] - \Omega_i[\epsilon(y_t)]\right]\right\}, \quad (3.4)$$

plus again the terms that refer to the feasibility constraint. The newly introduced variable, $M_i(y^{t-1})$, is defined recursively as $M_i(y_{-1}) = \lambda_i$, and

$$M_i(y^t) = M_i(y^{t-1}) + \mu_i(y^t). \quad (3.5)$$

Note that, at time t, the $M_i(y^t)$ are equal to the original weights plus the cumulative sum of the Lagrangian multipliers on the enforcement constraint at all periods from 1 to t. The first-order conditions that can be derived from this modified Lagrangian include

$$\frac{u'[c_1(y^t)]}{u'[c_2(y^t)]} = \frac{M_2(y^{t-1}) + \mu_2(y^t)}{M_1(y^{t-1}) + \mu_1(y^t)}, \quad (3.6)$$

in addition to the complementary slackness conditions. The next step consists of renormalizing the enforceability multipliers by defining

$$\varphi_i(y^t) = \frac{\mu_i(y^t)}{M_i(y^t)}, \qquad x(y^t) = \frac{M_2(y^t)}{M_1(y^t)}. \quad (3.7)$$

The virtue of this normalization is that it allows us to keep track of only the relative weight x. Its transition law can be written as

$$x(y^t) = \frac{[1 - \varphi_1(y^t)]}{[1 - \varphi_2(y^t)]}x(y^{t-1}), \quad (3.8)$$

by noting that $[1 - \varphi_1(y^t)]M(y^t) = M(y^{t-1})$.

We are now in a position to write this problem recursively. To do so, we define a mapping $\mathbf{T}$ from values into values, a fixed point of which are the value functions that characterize the solution to our problem. To solve our model numerically, as we do in the next section, we actually follow this procedure; that is, we iterate from a certain initial set of value functions. Successive approximations have yielded, in every case, the desired fixed point. The state variables are the current value of the shock y (recall that, because the shocks are Markov, their current value is sufficient to evaluate conditional expectations) and the current value of the relative weights x. Let $\mathbf{V} = \{V_0(y, x), V_1(y, x), V_2(y, x)\}$ be three functions, one for the planner and one for each of the agents, that satisfy

the following property:

$$V_0(y, x) = V_1(y, x) + x V_2(y, x). \tag{3.9}$$

The mapping $\mathbf{T}$, whose fixed point we are looking for, updates these three functions, and, therefore, we write the updated functions as

$$\mathbf{T}(\mathbf{V}) = \{T_0(\mathbf{V}), T_1(\mathbf{V}), T_2(\mathbf{V})\}.$$

To define $\mathbf{T}$, we first solve the following auxiliary problem where no incentive constraints are taken into account:

$$\Phi(y, x; \mathbf{V}) = \max_{c_1, c_2} u(c_1) + x u(c_2) + \beta \sum_{y'} \Gamma_{y,y'} V_0(y', x), \tag{3.10}$$

subject to the feasibility constraint (2.2), with solution $c_i^{\Phi,\mathbf{V}}$. Note that, in this problem, the relative weight x is constant. Next, we verify the enforceability of the solution to (3.10). This means verifying whether

$$u\left[c_i^{\Phi,\mathbf{V}}(y, x)\right] + \beta \sum_{y'} \Gamma_{y,y'} V_i(y', x) \geq \Omega[\epsilon(y)] \quad \text{for } i = 1, 2. \tag{3.11}$$

If (3.11) is satisfied, then $T_0(\mathbf{V}) = \Phi(y, x;\ \mathbf{V})$, and $T_1(\mathbf{V})$ and $T_2(\mathbf{V})$ are given by its left-hand side. It is easy to see that (3.11) cannot be violated for both agents at the same time (just note that autarky is a feasible allocation). The only remaining problem is to update the value functions when the constraint is binding for one of the agents, say agent 1. In this case, we solve the following system of equations in $\{c_1, c_2, x'\}$:

$$\Omega[\epsilon(y)] = u(c_1) + \beta \sum_{y'} \Gamma_{y,y'} V_1(y', x'), \tag{3.12}$$

$$x' = \frac{u'(c_1)}{u'(c_2)}, \tag{3.13}$$

$$c_1 + c_2 = e_1(y) + e_2(y) + 2\tau(y), \tag{3.14}$$

with solution $\{\bar{c}_1, \bar{c}_2, \bar{x}'\}$.[12] To update the value functions, we let

$$T_1(\mathbf{V})(y, x) = u(\bar{c}_1) + \beta \sum_{y'} \Gamma_{y,y'} V_1(y', \bar{x}'), \tag{3.15}$$

$$T_2(\mathbf{V})(y, x) = u(\bar{c}_2) + \beta \sum_{y'} \Gamma_{y,y'} V_2(y', \bar{x}'), \tag{3.16}$$

$$T_0(\mathbf{V})(y, x) = T_1(\mathbf{V})(y, x) + x T_2(\mathbf{V})(y, x). \tag{3.17}$$

A fixed point of $\mathbf{T}$, that is, $\mathbf{V}^* = \mathbf{T}(\mathbf{V}^*)$, gives a value to the problem of maximizing a weighted sum of utilities. Moreover, it also gives us a way to completely characterize the properties of such a solution by numerical methods. This means that, for any parameterization, we can tell whether the enforceable allocation

[12] There will typically be only one solution given the monotonicity of all the functions involved.

is autarky, the first best, or anything in between. We can also study how the enforceable allocations are affected by changes in the environment.

Note how different this type of problem is from a standard optimization problem. Note that there is more than one relevant set of first-order conditions: Binding states are represented by alternative Euler equations characterized by the default constraints.

3.2.2. Alternative Solution Methods

In the literature, alternative solution methods have also been proposed. Following the original, Thomas and Worrall (1988) and Ligon et al. (2000, 2001) characterize the present model by considering the set of Pareto-efficient allocations that also satisfies the participation constraints. This procedure leads them to characterize the solution to the problem in terms of a set of state-dependent intervals for the ratio of marginal utilities. The evolution of the ratio of marginal utilities is then described as follows. If, on one hand, the ratio of marginal utilities of consumption at time t follows within the interval of the state that occurs at date t', the ratio of marginal utilities is kept constant. If, on the other hand, the existing ratio of marginal utilities follows outside the interval, the program adjusts consumption so as to move the ratio of current marginal utilities within the intervals, but moving it by the smallest amount.

Ligon et al. (2001) stress how such a method highlights the fact that self-enforcing insurance contracts are midway between credit and insurance, a concept that is also stressed by Platteau (1997). Suppose, for instance, that we start from a situation in which the ratio of marginal utilities is one. In such a situation, first best would imply a simple sharing of the total output. Suppose now that individual 1 is relatively luckier and that, corresponding to that particular state of the world, the current ratio of marginal utilities follows outside from the relevant interval. This means that individual 1 is "constrained." The program implies that the transfer individual 1 makes is smaller than first best and moves the ratio of marginal utilities to the limit of the interval for that particular state; in this case, the relative weights in the planner function will move in favor of individual 1. If, in the following period, the shock is the same for the two individuals (and the ratio of marginal utility falls within the relevant interval), the ratio of marginal utilities will be kept constant. This implies transfes from individual 2 to individual 1. Notice that in first best these transfers would be equal to zero. What is happeining is that individual 2 is paying back individual 1 for his or her previous tranfser. When a new state of the world throws the marginal utility outside the relevant interval, the analogy with a credit contract ends. Indeed, the previous story is erased completely and the insurance aspect of the contract becomes more apparent.

There has been an attempt to use the mathematical apparatus developed by Abreu et al. (1990) with the value sets as arguments of the operators that they define directly as a means of computing equilibria. This is computationally very demanding, because it requires iteration in value sets (actually convex

sets), something that is very hard to do. Examples of this work include that by Judd and Conklin (1993) and Phelan and Stacchetti (2001). These two papers use different algorithms to store sets, and, to our knowledge, the methods that they have developed have not been used by other researchers.

Alvarez and Jermann (1998) have a very interesting and useful way to characterize aggregate growth and aggregate uncertainty in models with imperfect enforceability. They show that, for these phenomena to be considered, it is sufficient to redefine the discount factor (and make it state and time dependent in a particular way). Moreover, they present calibrations that show that the possibility of default makes asset pricing more similar to observed data. In particular, Alvarez and Jermann (1998) go some way toward explaining the equity premium puzzle and the low values of real interest rates on safe assets. They stress the similarities between their approach and the recent study by Luttmer (1999).[13]

3.2.3. Extensions and Complications

The simple model we just presented can be extended and complicated in a variety of directions. First, one can consider many households, rather than only two. Results with many households are presented by Ligon et al. (2000, 2001) and by Alvarez and Jermann (1998). The analysis does not present any conceptual difficulty. The central planner problem will have to be modified for the participation constraints of all the households involved to be considered. In terms of the approach by Ligon et al. (2000), one has to consider a multidimensional Pareto frontier.

A more complex extension is the consideration of storage possibilities. Storage is difficult for several reasons. First, the solution of the problem has to determine where the investment takes place. When the rate of return is independent of the size of the investment, the solution of the first best is that it is irrelevant where the investment takes place. When contracts have to be self-enforceable, the location of the investment affects the conditions in which each agent reverts to autarky, posing an additional margin to induce participation in the scheme that has to be taken into account. This problem is avoided by Ligon et al. (2000) by modeling a technology in which the process of reversion to autarky involves the loss of assets whereas being in autarky is compatible with savings. They defend this assumption on the basis of the possibility that storage is, in any case, a communal thing (maybe administered by a local authority). In this case, deviants will not have the possibility of storing under

[13] Alvarez and Jermann (1998) use yet a different characterization of this type of model. As they are interested in the asset-pricing implications of this type of model, they consider contingent loans. The lack of enforceability is then reflected in the impossibility, for individuals affected by negative shocks, to borrow more than an amount over which they would have no incentive to default in any state of the world next period. Although the characterization of the equilibrium is somewhat different, clearly the results are very similar.

autarky. Second, the presence of storage introduces nonconvexities into the problem that make the numerical solution of the problem particularly complex.

In a recent paper, Lambertini (1999) considered a three-period overlapping generations model with limited enforceability. In such a model, one has to assume the presence of storage possibilities. At least three periods are necessary; otherwise, no contract could ever be enforced. Moreover, autarky is defined as a situation in which people not only cannot borrow, but also are prohibited from saving (in that their savings can be appropriated). This framework allows Lambertini to consider the effects of limited enforceability in a finite-lives context. In particular, Lambertini shows that, if the income profile is hump shaped, then the equilibria generate borrowing constraints for young individuals that prevent consumption smoothing. In the absence of commitment, the model, in general, has multiple equilibria.

Kehoe and Levine (1993, 2001) also consider models with limited commitment. In these, agents cannot borrow as much as they could with perfect commitment, because there is always the possibility of bankruptcy. This possibility induces an endogenous borrowing limit.

3.3. Empirical Evidence

Although many papers have looked at the determinants of private transfers and at their interaction with public transfers (see what follows), the evidence on imperfect enforceable models is limited. To our knowledge, only a handful of papers have taken the models we discussed herein to the data in order to test their implications. We start by reviewing some evidence on the crowding out of private transfers by public transfers. We then consider the little empirical evidence on models with imperfect enforceability and conclude with possible extensions.

3.3.1. Transfers and Crowding Out

Several papers have analyzed private transfers among families and how these interact with the provision of public transfer programs. Obviously, models with imperfect enforceability are not necessarily the only ones that can be used to analyze the interaction between private and public ones, especially if the latter do not have an insurance component. It is possible, for instance, that transfers are motivated by altruism.[14] In such a framework, public transfers will certainly crowd out private transfers, under standard assumptions on preferences. More generally, it is possible what we measure as transfers is given in exchange for some sort of service. For instance, children might transfer resources to parents in exchange for help with small kids, or parents might transfer resources to children in exchange for support and care. Indeed, the model we discussed herein is a particular kind of exchange, as current transfers are given in order to receive

[14] See Cox (1987a).

future transfers. Although in the model we consider, in which the effect of public transfers works through a reduction of the variance of aggregate shocks, the effect of public transfers is unambiguously negative, there are other situations in which such an effect can be ambiguous. Examples of these situations are given in Cox (1987b).

Several studies have analyzed the extent to which public transfers of a different nature, ranging from social security and pensions to food aid, crowd out private transfers. Cox and Jakubson (1995) analyze U.S. data on Aid to Families with Dependent Children, whereas Cox, Eser, and Jimenez (1998) look at whether private income transfers are affected by public transfers in Peru.[15] Jensen (1999) analyzes the relationship between migrant remittances and the possible recipients of old-age pension in South Africa. In most of these studies, the hypothesis that private transfers are crowded out by public ones is tested by means of simple regressions in which the dependent variables are private (net or gross) transfers or remittances and the independent variables include, in addition to standard controls, some indicators of whether the household receives some form of public transfers (e.g., public pensions). Some studies look at the intensive margin, whereas others consider the extensive margin or both. Typically, either probit or tobit models are used, even though recently some researchers have also tried nonparametric methods.[16]

In addition to the nonlinearity of models where the dependent variables are discrete or truncated, the main problem faced by empirical researchers in this area is that of endogeneity. Typically, the beneficiaries of public transfer schemes are not chosen randomly, creating, in all likelihood, an important endogeneity problem. It is therefore particularly valuable to consider the effect that public transfers have on private transfers when there is a good "instrumental" variable available. Albarran and Attanasio (2000) expand the simple exercise performed in Attanasio and Ríos-Rull (2000a) and consider the effect that a large public transfer program, called Progresa, has had in rural Mexico. The Mexican data set is particularly attractive, because, when the program was started, the decision was made to evaluate it. For such a purpose, a number of villages were randomized out of the program for two years. One can therefore compare beneficiaries in the treatment villages and would-be beneficiaries in the control villages.

The results obtained in Albarran and Attanasio (2000) indicate a substantial amount of crowding out. Such a result holds both when a probit is looked at for any kind of transfer (monetary or in kind) and when a tobit is looked at for monetary transfers.

[15] In particular, Cox et al. (1998) find that public transfers have a positive effect on the amount of private transfers, but that "social security benefits crowd out the incidence of private transfers."

[16] See, for instance, Jensen (1999). The use of nonparametric methods to estimate truncated models implies some serious identification problems. In particular, identification requires that one find a variable that affects the extensive margin (whether the household receives a transfer or not) that does not affect the quantity of transfer received. This is obviously a difficult problem.

As we discuss in Section 4, a transfer that moves the mean of the endowment causes a decrease in the level of private transfers. The results obtained by Albarran and Attanasio (2000), therefore, are consistent with the implications of the model with imperfect enforceability.

3.3.2. Evidence From Models With Imperfect Enforceability

To the best of our knowledge, the only two papers with direct evidence from these models are those by Foster and Rosenzweig (1999) and Ligon et al. (2000). In addition, Attanasio and Ríos-Rull (2000a) have interpreted some of the evidence from the Progresa program in Mexico as relevant for the importance of the incentives created by imperfect enforceability. Krueger and Perri (1999) have interpreted some U.S. evidence on consumption and income inequality as consistent with the type of models we discussed. Alvarez and Jermann (1998) have provided some calibration in favor of these models. Finally, Platteau and Abraham (1987) explicitly refer to models with imperfect enforceability to explain their evidence, whereas the evidence from Udry (1994) and Fafchamps and Lund (2000) is consistent with these models.

The paper by Foster and Rosenzweig (1999) is the only one that uses evidence on transfers to assess the relevance of the models we have been discussing. Moreover, they extend the model to incorporate an altruistic motive in individual preferences. Rather than estimating a fully structural model, Foster and Rosenzweig notice that the absence of perfect enforceability and an endogenous missing market has specific implications for the time-series properties of transfers. In particular, they note that, conditional on current shocks, current transfers should be negatively related to the cumulative amount of past transfers. First differencing such an equation, they obtain a relationship that can be estimated in panel data. In particular, they find that the change in transfers should be inversely related to the lagged level of transfers, once one conditions on shocks. Given that it is difficult to obtain a closed-form solution for transfers in such a model, to have an idea of what a plausible coefficient of such a regression is, Foster and Rosenzweig run similar regressions on data generated by simulating a model like those we discussed. They find evidence that the regression coefficient obtained from simulated data is not too different from those obtained on actual data from Pakistan and India.

Ligon et al. (2000), instead, take a fully structural approach. They consider a model very much like the one we considered herein, and, by matching it to the ICRISAT data (originally used by Townsend, 1994, to test the implications of perfect insurance), they estimate the structural parameters of the model by maximum likelihood. The computation of the likelihood function is numerically very intensive, as it involves computing, for each value of the parameters, the decision choices implied by the model. For this reason, Ligon et al. (2000) are forced to drastic simplifications. Even though their theoretical model can account for storage and for multiple agents among whom the insurance contracts are stipulated, in the empirical applications they assume that there is no

storage and that each household plays a game with the rest of the village, rather than considering all the households at the same time. Even with these strong simplifications, Ligon et al. (2000) find that the model fits the data remarkably well and much better than both the perfect insurance model and the static model considered by Coate and Ravallion (1993). In particular, Ligon et al. (2000) show that the correlation between actual consumption and consumption generated by the model is highest for the model with imperfect enforceability (over perfect insurance, autarky, and the Coate and Ravallion, 1993, model).

All the papers we mentioned are not exempt from criticism. The most crucial point is the fact that these models neglect to consider storage. As we already discussed, the possibility of storage is unlikely to affect the *qualitative* features of the equilibrium. However, storage directly affects the amount of risk sharing that can be sustained in equilibrium. Therefore, an empirical exercise that neglects the possibility of storage might lead to results that are seriously biased, especially in situations in which storage is, at some level, an important way to smooth out idiosyncratic shocks.

3.3.3. Extensions

The existing papers are important first steps in the study of the implications of models with imperfect enforceability. However, there is still much work that has to be done. In our opinion, there are two particularly fruitful directions this research can take, even though the data requirements to implement these ideas can be quite formidable.

In his paper, Kocherlakota (1996) devotes the last section to the discussion of the empirical implications of models with imperfect enforceability. Kocherlakota stresses how the empirical implications of models with imperfect enforceability can be different from those of models with imperfect information. He notes that the model can be rewritten as one in which the weights of the social planner problem change with the shocks received by the individual agents. As the weights that determine the allocation of resources are equal to the ratio of marginal utilities at the beginning of the period, a testable implication of the model is that, conditional on the lagged ratio of marginal utilities, current consumption should not depend on any other lagged information. In private information situations, this is not necessarily the case. Although this is an interesting result, Kocherlakota stresses that this is not a strong implication. The fact that the ratio of marginal utilities is a sufficient statistic for current consumption does not necessarily imply constrained efficiency of the resource allocation, which is what the models we consider imply.

However, Kocherlakota notices that, for people that are unconstrained, in that their participation constraint is not binding, there is a standard Euler equation that governs the evolution of consumption. Moreover, only a subset of individuals in a village can be constrained in a certain period; that is, the participation constraint cannot be binding for everybody. In addition, if we find an unconstrained individual, we can be sure that individuals with a higher marginal utility

of consumption are also unconstrained. The model then has strong implications for constrained and unconstrained individuals. For the unconstrained, the ratio of marginal utilities will be constant and completely determine consumption. For the constrained, instead, own marginal utility will not be relevant for the determination of current consumption, which will instead be determined by the value of shocks and by the marginal utility of unconstrained individuals.

Kocherlakota's characterization of the implications of the model is a very interesting one. However, it is also clear that the data requirements are quite formidable. One would want to use time-series data covering a long-enough period to warrant precise estimates. Notice that, as with the estimation of Euler equations for consumption, consistency requires "long-T" asymptotics. Alternatively, one could use information about several villages if one was willing to assume that the shocks received by the villages were independent and that the villages were isolated.

Ligon et al. (2000) also consider the Euler equation that links the equilibrium evolution of consumption. This can be written as follows:

$$u_i'(c_s^i) = \beta(1+R)E_s\left[u_i'(c_r^i)\right] + \omega_i/\lambda_i + \beta E_s\left\{\phi_r^i\left[(1+R)u_i'(c_r^i) - f_r'^i\right]\right\}, \tag{3.18}$$

where β is the discount factor, R is the fixed and exogenous return on the storage technology (that can be negative), λ_i is the initial Pareto weight in the social planner problem, ω_i is the multiplier on the constraint that storage cannot be negative, ϕ_i is the multiplier on the participation constraint, and f' represents the marginal effect of an additional unit of storage on the value of autarky. Ligon et al. (2000) provide a nice interpretation of Equation (3.18). The first term represents the traditional effect that a reduction in current consumption will translate, through the storage technology and the discount factor in future utility. The second term also takes into account the effect that standard liquidity constraints have on consumption. The third effect captures the effect that an additional unit of saving has on the participation constraints. Whether this is positive or negative depends on which of the two effects in the square brackets prevails. The last term depends on the particular arrangements for storage. If storage is held comunnally, it is zero. Notice that there might be some individuals for whom both ω_i and ϕ_i are equal to zero. These are individuals who are sufficiently "lucky" so as to wish to be saving, but not "lucky enough" for their participation constraint to be binding. On the two sides of these individuals, there are those for whom ϕ_i is positive and those for whom ω_i is positive. Note that both multipliers cannot be positive at the same time because agents will be tempted to go to autarky only when they have stored a sufficiently large amount of the good, not when they would like to move resources from the future to the present.

This type of equation has yet to be exploited. There are two types of difficulty. The first is common to the implementation of the tests proposed by Kocherlakota, that is, the fact that the data requirements, especially in terms of

the length of the period, can be prohibitive. The other is the fact that the multipliers present in Equation (3.18) are not observable. One possibility would be to parameterize the Kuhn–Tucker multipliers and relate them to past income shocks, as well as to other observable characteristics.

4. IMPERFECT ENFORCEABILITY: CHARACTERIZING EQUILIBRIA

The aim of this section is to characterize the properties of the equilibria of the models with imperfect enforceability discussed herein. In particular, we are interested in establishing the relationship between changes in various preference and technology parameters and the amount of risk sharing that can be sustained in equilibrium when contracts are not perfectly enforceable. For many of the parameters of the model, analytical results are available. For instance, it is obvious that an increase in the discount factor β induces more risk sharing. In other cases, however, it is necessary to use numerical simulations to characterize the equilibria, and, even when the effect of changes in the parameters of the model can be signed analytically, it is important to quantify these effects in various situations. It is for this reason that, while mentioning some analytical results, we focus on the results of numerical simulations.

In the numerical simulations, given the parameters of the model, we solve for the consumpion functions in the environment just described and use them to simulate the model for 40,000 periods. We discard the first 100 periods and compute the relevant statistics, averaging across simulations. All the simulations are generated with the same initial seed, so that the realized sequence of shocks is the same for all economies.

The parameters of the various experiments we perform are presented in the top panel of Table 6.1. These include the preference parameters (risk aversion and discount factor) and the parameters that characterize the endowment processes (points of support and transition matrices). In this part of the table we also report the implied properties of the income processes, such as its mean, standard deviation, and first-order autocorrelation.

In the bottom panel of Table 6.1, we report the standard deviation and autocorrelation of consumption in the three types of equilibria (enforceable, first best, and autarky). Finally, we report three statistics that are indicative of the amount of risk sharing that is achieved in the equilibrium with self-enforceable contracts: the average level of private transfers as a percentage of per capita income and as a percentage of the transfers in the first-best equilibrium and, finally, the fraction of consumption cross-sectional variance (after removing aggregate consumption) over the variance of endowments. Such a ratio should be zero if first best is achieved, and one under autarky.

We start our discussion by looking at a baseline specification, whose features are reported in the first column of Table 6.1. The discount factor is equal to 0.85, and the coefficient of relative risk aversion is equal to 1.1. Both aggregate and idiosyncratic shocks can take two values. Bad shocks are quite severe: the low value for both aggregate and idiosyncratic shocks is equal to 10 percent

Table 6.1. *Properties of baseline and other specifications*

	(1) Base line	(2) $+\beta$	(3) $+\gamma$	(4) + Id Per	(5) − Ag Var	(6) + Ag Per b	(7) + Ag Per g	(8) + Ag Inc
				Preference Parameters				
β	0.85	**0.88**	0.85	0.85	0.85	0.85	0.85	0.85
γ	1.1	1.1	**1.3**	1.1	1.1	1.1	1.1	1.1
				Idiosyncratic Shock Process				
Values	1,0.1	1,0.1	1,0.1	1,0.1	1,0.1	1,0.1	1,0.1	1,0.1
$\Gamma_{g,g}$, $\Gamma_{b,b}$	[0.7,0.7]	[0.7,0.7]	[0.7,0.7]	[**0.8,0.8**]	[0.7,0.7]	[0.7,0.7]	[0.7,0.7]	[0.7,0.7]
				Aggregate Shock Process				
Values	1,0.1	1,0.1	1,0.1	1,0.1	**0.98,0.28**	**1.01,0.21**	**0.99,0.01**	**1.2,0.3**
$\Gamma_{g,g}$, $\Gamma_{b,b}$	[0.9,0.1]	[0.9,0.1]	[0.9,0.1]	[0.9,0.1]	[0.9,0.1]	[0.9,**0.3**]	[**0.92**,0.1]	[0.9,0.1]
				Output Statistics				
Aver.	1.46	1.46	1.46	1.46	1.46	1.46	1.46	1.66
St. dev.	0.52	0.52	0.52	0.52	**0.50**	0.52	0.52	0.52
Autocorr.	0.29	0.29	0.29	**0.44**	**0.33**	**0.35**	**0.30**	0.29
Aggr.	0.00	0.00	0.00	0.00	0.00	0.20	0.01	0.00
Idios.	0.40	0.40	0.40	0.40	0.40	0.40	0.40	0.40
				Properties of the Allocations				
				Enforceable Consumption				
St. dev.	0.45	0.43	0.44	0.48	0.43	0.45	0.44	0.47
Autocorr.	0.29	0.27	0.30	0.43	0.34	0.37	0.30	0.30
				First-Best Consumption				
St. dev.	0.42	0.42	0.42	0.42	0.38	0.42	0.42	0.42
Autocorr.	0.23	0.24	0.24	0.35	0.28	0.32	0.25	0.23
				Autarkic Consumption				
St. dev.	0.52	0.52	0.52	0.53	0.50	0.52	0.52	0.52
Autocorr.	0.29	0.29	0.29	0.44	0.33	0.35	0.30	0.30
				Average Enforceable Transfer as % (of)				
Income	0.138	0.148	0.048	0.077	0.097	0.119	0.144	0.065
1st best Tr.	0.905	0.966	0.952	0.505	0.627	0.771	0.936	0.478
				Var. of Consumption/Var. of Endowments				
	0.271	0.139	0.152	0.537	0.377	0.310	0.221	0.490

Note: Aver. = average, st. dev. = standard deviation, Autocorr = autocorrelation, Var. = variance, Aggr. = aggregate, Idios = idiosyncratic, 1st best Tr. = 1st-best transfers.

of the high value. The persistence properties of the two shocks, however, are quite different. Bad aggregate shocks are much rarer and less persistent than idiosyncratic shocks. Overall, the first-order autocorrelation of aggregate shocks is close to 0, whereas that of idiosyncratic shocks is about 0.4. The overall autocorrelation of per capita output is 0.29.

The utility achieved by self-enforceable contracts is in between that under autarky and that under first best. In this particular example, the intertemporal allocation of resources is closer to first best than to autarky. This can be seen

by looking at the ratio of the cross-sectional variance of consumption to the cross-sectional variance of endowments (last row of Table 6.1): the value of 0.27 is closer to 0 than to 1. The amount of risk sharing that happens is also reflected in the relatively high (absolute) value of private transfers, which, on average, are equal to 14 percent of output (and 90 percent of the transfers that would occur in first best). Also notice that the standard deviation of consumption of the enforceable allocation is midway between the standard deviations of consumption in first best and in autarky. The autocorrelation of consumption under the incentive compatible equilibrium is not different from the autocorrelation that would be observed under autarky and substantially higher than that which one would observe under first best. Indeed, it is possible to have cases in which the constrained efficient equilibrium generates more persistence in consumption than in autarky.

The relatively high persistence of consumption in the constrained efficient equilibrium is related to another interesting phenomenon. As we just mentioned, it is possible that, in this equilibrium, transfers are not only smaller but of the opposite sign than those that would be observed in the first best. Even in the simple example we propose here, this does happen. For the parameters in the first column of Table 6.1, on average, the system finds itself 3.5 percent of the time in a situation in which private transfers flow from the relatively poorer to the relatively richer individual. This happens only when the aggregate state is good and the ratio of marginal utilities (the state variable in our recursive formulation) has reached a certain trigger level. This happens when one of the two individuals has been relatively "lucky" for some periods so that his or her participation constraint has been binding repeatedly.

In columns (2)–(8) we change various parameters of the system. In particular, in columns (2) and (3) we increase the values of the two preference parameters, whereas in columns (4)–(8) we change the parameters of the endowment processes. The parameters that change relative to the baseline specification are in boldface in Table 6.1.

In column (2) we increase the rate of discount from 0.85 to 0.88. As expected, this has the effect of increasing risk sharing and bringing the constrained efficient equilibrium closer to first best and farther away from autarky. The same happens when we increase the coefficient of relative risk aversion, which is done in column (3). In this column, γ is changed from 1.1 to 1.3. Once again, we see an increase in the amount of risk sharing. The ratio of the variance of consumption to the variance of endowments moves from 0.27 to 0.15, whereas the transfers as a fraction of first-best transfers increase from 0.91 to 0.95. Notice that, even though the changes in the parameter values are quite small, their effects are sizeable.

In column (4), we start looking at changes in the economic environment in which the agents live. First, we increase the persistence of idiosyncratic shocks. This has the unambiguous effect of reducing risk sharing, for reasons that should be, by now, obvious. An agent receiving a very persistent and positive shock will not want to share it (or will not want to share a persistent negative shock received by his or her partner).

In column (5), we decrease the variance of the aggregate shocks, leaving the mean (and the other income processes) unaffected. This exercise is relevant for evaluating the provision of aggregate insurance by international organizations. In the resulting enforceable equilibrium, there is substantially less risk sharing than in the baseline case. Private transfers as a ratio of first-best transfers decline from 0.91 to 0.64, whereas the ratio of the variance of consumption to the variance of endowments increases from 0.27 to 0.36. Notice that this is a case in which the persistence of consumption in the enforceable equilibrium is larger than the persistence in autarky.

In columns (6) and (7), we increase the persistence of aggregate shocks, leaving their mean and variance unaffected. We do so in column (6) by increasing solely $\Gamma_{b,b}$, the probability of the bad shock repeating, while leaving $\Gamma_{g,g}$, the probability of the good shock repeating, unaltered. Note that for the mean and standard deviation of output to be the same as in the baseline economy, the values of the shocks in both states have to be higher, which implies that the bad state is less painful, even if more frequent. It turns out that the increase in $\Gamma_{b,b}$ causes a decrease in risk sharing as measured both by the fraction that average enforceable transfers represent of first-best transfers (goes down to 0.771 from 0.905) and by the ratio of the variances of consumption and endowments (goes up to 0.310 from 0.271). The reason for this is likely to be related to the issue that aggregate bad times are better and more likely to last than in the baseline, making autarky less painful, which reduces the gain from cooperation.

In column (7) it is $\Gamma_{g,g}$ that we increase, leaving $\Gamma_{b,b}$ unchanged to increase the autocorrelation of the aggregate shock. We again adjust the values of the shock to leave the mean and standard deviation unchanged. This adjustment requires that the values of both states of the aggregate shock are lower, even if the bad state is less frequent. In this case the enforceable allocation gets closer to the first best relative to the baseline. The reason, like in the previous economy, has to do with the value of autarky. In this economy, autarky is more painful because the bad states are worse than in the baseline. As a result, agents are more willing to cooperate.

In column (8), we increase the mean of the aggregate process by 0.2, leaving its variance unaffected. One can think of this as a subsidy that gets distributed to everybody in the village. The effect of such a scheme, which obviously increases welfare, is to reduce the amount of risk sharing that can be sustained in equilibrium. By moving the system away from zero, the scheme moves individuals away from states with really high marginal utilities of consumption. This means that the punishment implicit in autarky is not as harsh as in the baseline. Therefore there will be less risk sharing and private transfers. Notice that an increase in the mean of aggregate endowment that would also increase the variance but leave the coefficient of variation of the endowment process unchanged (such as a multiplicative shift) would leave the amount of risk sharing unaffected. This is, however, a consequence of the assumption of homothetic preferences. With a Stone–Geary utility function, an increase in the mean induced by a multiplicative shift would also reduce risk sharing.

5. THE PROVISION OF AGGREGATE INSURANCE UNDER IMPERFECT ENFORCEABILITY

Here we consider the possibility that an outside agent with taxing powers, such as the government or an international organization (the World Bank), offers insurance against the aggregate shocks, which is an opportunity that the households within the village cannot afford but that can be provided by an external entity. There are two important points we want to make in this section. First, we want to stress that the provision of aggregate insurance, like most government interventions, does not happen in a vacuum. In all likelihood, the provision of aggregate insurance interacts with the functioning of private markets. In Subsection 5.1, we describe what happens to the amount of risk sharing that occurs in equilibrium when, in the presence of enforceability problems, the government introduces an insurance scheme that smooths out part of the aggregate fluctuations. We call a reduction in the amount of risk sharing following the introduction of such a scheme "crowding out." We also show, within the framework of the model proposed herein, that it is possible that the provision of aggregate insurance may make individual agents worse off. Such a situation occurs when the crowding out induced by the policy more than offsets the benefits of the insurance that the public policy provides.

In Subsection 5.2, we make the point that, given the effect that the provision of aggregate insurance has on the functioning of private markets and incentives, it is worth thinking about the design of such schemes carefully. In particular, we ask whether it is possible to introduce aggregate insurance schemes that avoid the crowding out of idiosyncratic insurance. We show that, in general, the answer to such a question is yes.

5.1. Crowding Out Results

In the model herein, we can introduce aggregate insurance, provided by an external agent, in a very simple fashion. We consider transfers that are contingent on the aggregate shock only. Suppose, for simplicity, that the aggregate state can take only two values, low and high. We assume that the central government collects a premium in good aggregate states and pays out the actuarially fair amount corresponding to such a premium in aggregate bad states.[17]

The effects of such a scheme are immedetiately apparent in the model we just considered, in that they will enter both the value of autarky and the continuation value of insurable contracts. As the effect of the government scheme is equivalent to a reduction in the variance of the aggregate shock, we know already, from Table 6.1, that the effect of such a scheme will be a reduction in

[17] One way to think about such a scheme is that the government is smoothing shocks across villages by maintaining a balanced budget at each point in time. Alternatively, we could think of the possibility that the government has access to a storage technology that is not available to the individual agents.

the amount of risk sharing that occurs in equilibrium. In particular, we would have a reduction in the average size of private transfers and an increase in the ratio of the variance of consumption (net of aggregate consumption) over the variance of endowments.

Moreover, it is possible to construct examples in which the introduction of a mandatory aggregate insurance scheme makes agents worse off. Such an example is provided in Attanasio and Ríos-Rull (2000a). It may be expected that a welfare decrease would occur if we start from a situation in which there is a substantial amount of risk sharing that goes on and that can be crowded out. This is not the case in the example in Attanasio and Ríos-Rull, in which shocks are very extreme and idiosyncratic shocks very persistent. In such a situation, high variance induces high risk sharing, whereas high persistence induces low risk sharing. It turns out that for the values used by Attanasio and Ríos-Rull, there is very little risk sharing: The ratio of variances is 0.85. However, the little risk sharing that goes on happens at crucial moments, when things are bad at both the aggregate and the idiosyncratic levels. In such a situation, the crowding out of a little insurance does make people worse off.

The result that the introduction of aggregate insurance can lead to a decrease in welfare is reminiscent of a similar result derived by Ligon et al. (2001), who show that the introduction of a storage technology in an economy with imperfect enforceability and no storage can lead, under certain circumstances, to a welfare decrease. The reason why this can happen is the same as the reason why the introduction of aggregate insurance might decrease welfare, that is, the increase in the value of autarky and the subsequent decline in idiosyncratic risk sharing. Another relevant result is the one recently derived by Krueger and Perri (1999), who, in a model with partly insurable idiosyncratic shocks but no aggregate shocks, showed that progressive taxes can reduce the amount of risk sharing by increasing the value of autarky.[18]

5.2. On the Optimal Design of Aggregate Insurance

Given the results discussed in Subsection 5.1, it makes sense to ask whether it is possible to design the aggregate insurance scheme to prevent or minimize the problems we discussed. In this subsection, we explore this possibility using material from Attanasio and Ríos-Rull (2000b).

In the model we just used, the crowding out of idiosyncratic insurance originated from the fact that the aggregate insurance scheme increases the value of autarky and therefore decreases the individual incentives agents have to insure each other. In the absence of enforcement mechanisms, the equilibrium is sustained by the threat of depriving individual agents of future smoothing mechanisms. However, in the model we described, individual agents can deny their partners only the insurance of idiosyncratic shocks.

[18] Di Tella and MacCulloch (1999) stressed similar points when they considered a welfare system.

A possibility worth exploring, therefore, is to create a scheme that gives individuals the possibility of punishing their partners by denying them not only the idiosyncratic insurance, but also the aggregate one. Here we consider two alternative schemes that aim at achieving this result. The main idea is to make individuals play games that, in equilibrium, will yield an allocation of resources that is sustained by off-equilibrium payoffs that imply the disappearance of the aggregate insurance. For this reason, such schemes avoid the crowding out of private insurance.

Some readers will have noticed the similarity of the previous paragraph to the language of implementation theory.[19] Indeed, our scheme uses some implementation theory ideas. However, there are also important differences. Much (but not all) of this literature is quite abstract, in the sense that, although it proves implementability of a given outcome in a certain situation, it does not necessarily specify the mechanism through which a given desired outcome is implemented. Furthermore, sometimes these mechanisms can be quite complex and esoteric.[20] We are not particularly interested in showing the possibility of implementing first-best allocations by some abstract and possibly very complicated mechanisms. Instead, in designing our games, we are guided by the principle that we want the schemes to be carried out in practice, possibly by a relatively unsophisticated bureaucracy. This means that we weigh simplicity over arbitrary closeness to the first best. Even though we do not explicitly model the workings of the bureaucracies of developing countries, we have their limitations very much in mind.

In our discussion, we focus on purposely simple and specific mechanisms and explore their ability to avoid some of the problems discussed in the previous section. These are not necessarily the optimal mechanisms and do not necessarily avoid the problem completely, as we shall see.[21] Moreover, in our mechanism, we do not use the type of mechanisms suggested in the implementation literature for the case in which there are at least three individuals. It should be stressed, however, that our "implementation" problem is simplified by the fact that the games we construct are not zero sum: We can use the utility provided by the smoothing of aggregate shocks (which is not available in the absence of the aggregate scheme we construct) as a possible carrot and stick to implement the desired outcomes.

As we have said, we explore two different schemes. Each of these schemes involves making the agents play a simple game in each period when the aggregate state is bad. The payouts the World Bank gives depend on how the agents play this game. The first scheme that we explore is very simple, but the equilibrium that we call for and that leaves the value of autarky unaffected does

[19] Moore (1992) provides a very entertaining and useful survey of the literature. Jackson (2000) provides a more recent and excellent survey.

[20] An important set of results state that when there are more than three agents, many outcomes are easily implementable (see Moore and Repullo, 1988).

[21] Certainly, they do not necessarily achieve the first-best allocation.

not dominate other equilibria in which the value of autarky changes with the introduction of the scheme in the same way as described herein. Indeed, treated as a one-shot game, such an equilibrium (which, as we discuss later, involves both agents playing "no") does not look very convincing, as it is dominated by another Nash equilibrium (yes, yes); neither equilibrium is strict (i.e., such that your payoff strictly falls should you unilaterally deviate from it), and indeed (yes, no) and (no, yes) are also equilbria.

The second scheme, instead, although slightly more complex, avoids some, but not all, of these problems by providing stronger incentives to implement punishment off the equilibrium and, at the same time, changes the value of autarky very little, making it more robust to the problem of the agents coordinating on an unattractive equilibrium. It consists of adding a small reward for saying no. This breaks the tie, and it makes the one-shot game a coordination game.

5.2.1. A Simple Scheme

The first scheme we propose is very simple. We require, on the part of the institution providing the aggregate insurance, knowledge of the typical size of the extended families.[22] Therefore, in the model we presented herein, we assume that such an institution knows the fact that extended families are formed of two individuals. When introducing the scheme, individuals are requested to register as pairs. In good times they will pay, as before, a premium, and in bad times they might be entitled to a payment that is the actuarially fair value of the premium paid. However, such a payment is not automatic, but is conditional on the outcome of a game. Such a game consists of asking, simultaneously and separately, the two agents in a pair whether the payment to the other individual should be executed. That is, whether an individual receives his or her payment or not depends on what the other individual says.

In the following matrix, Table 6.2, we describe the payoff structure of the game that is played every time a bad aggregate shock strikes the village. Note that it is the type of game that Moulin (1986) calls *give your friend a favor*, because your actions do not affect your own payoff, only the other players'. In our case, we use it as a gate to do the opposite and give your relative a punishment.

As is evident, such a mechanism gives each agent the possibility of punishing his or her partner, when and if the relationship breaks down, by depriving the partner not only of the idiosyncratic transfer but also of the aggregate payment. It is easy to construct equilibria for the repeated game between the two players of the following type. Say yes if the other agent has collaborated (given the transfer required by the contract), and play no otherwise. This is, of course, not

[22] It is likely, though, that the institution may offer a menu of deals, depending on the size of the families that show up at registration time. We leave for future research the issue of coexistence of family sizes.

Table 6.2. *First scheme: the simple-scheme payoff structure*

	2 says yes	2 says no
1 says yes	$\{P, P\}$	$\{0, P\}$
1 says no	$\{P, 0\}$	$\{0, 0\}$

the unique equilibria. In fact, if ever a player deviates, playing no is not even a strictly dominant strategy given that the other player says no.

Notice that the equilibrium that we describe has the feature that the stick the agents face when they deviate is unaffected by the government's policy, except, possibly, for the premium they will be paying in aggregate good times. This scheme, therefore, not only will avoid the crowding out, but will induce some crowding in. The reason for this is that, by providing aggregate insurance conditional on the relationship's lasting, we increase the value of being in the contract while at the same time lowering the value of autarky, because of the presence of the premium in good times, which, in autarky, will not be compensated for by a payment in bad times if agents play {no, no}. This causes the amount of individual risk sharing to increase. For instance, if the World Bank introduces an aggregate insurance scheme that reduces the variance of aggregate shocks by an amount equivalent to the reduction in aggregate variance discussed in column 5 of Table 6.1, the ratio of the variance of consumption to endowments (which is inversely related to the amount of risk sharing that is sustainable in equilibrium), which increased from 0.27 in column 1 to 0.38 in column 5, is actually reduced to almost zero when the scheme is supplemented by the simple game we propose here. In addition, it should be stressed that not all the job is done by the reduction in the value of autarky induced by the tax in aggregate good times. Even if we assume that the government disappears once the first no, no is played (so that the value of autarky is effectively not affected by the scheme), the ratio of variances goes down to 0.026, indicating that allocations very similar to first best can be achieved in the new equilibrium.

This result is important because it shows that the government or the relevant international institution can kill two birds with one stone. On one hand, it can smooth shocks across villages. On the other hand, at the same time, by carefully constructing the aggregate insurance scheme, it can use the extra utility it provides as a discipline device that improves the functioning of the private insurance market.

Note that the scheme we propose does not require the managing institution to have knowledge of who the members of the extended family are. In our framework, the members of the family will have an incentive to register as a pair, so as to precommit to an ex ante better equilibrium.

Given the "crowding in" result, it is legitimate to ask whether such a scheme can achieve first best. The answer is, in general, no. Obviously, if the variance of the aggregate shock is sufficiently high, the amount at stake becomes large

enough to make the first-best allocation self-enforcing. This is, however, a limit case. Obviously, as is the case in most repeated interactions, a folk theorem can be proved. Such a theorem states that, for sufficiently high discount rates, the first best can be achieved. When we state that the first best cannot be achieved, in general we take the standard position in economics that discount rates are part of the environment and cannot be manipulated by the researcher.

The main conceptual problem with this scheme is that, off the equilibrium, each agent will not have a strong incentive to punish his or her partner by denying the partner the aggregate payment. In particular, as such an action does not strictly dominate the alternative of allowing the partner to collect the payment, there are other equilibria in addition to the desirable one, in which the two partners never deny each other the aggregate payment in bad aggregate states. This implies that the value of autarky will change with the introduction of the scheme in the same way as it would change if the scheme was introduced without the game. To avoid these unicity problems, we now turn to a more complex scheme.

5.2.2. A Scheme That Rewards the Naysayers

Our second scheme works as follows. As in the simpler case described earlier, to participate in the scheme, individuals are asked to register as a pair. However, in bad aggregate states, the game that the two agents, labeled 1 and 2, are asked to play is described by the matrix shown in Table 6.3.

The entries in the four cells describe the monetary payoffs the two agents receive (depending on which both of them play) in the period they play the game. As before, P is the actuarially fair payment of the aggregate insurance scheme. However, e is a small additional transfer to those that say no; e is considerably smaller than P. In this scheme, each agent has the possibility of getting, in addition to the basic payment P, a transfer e. In getting the transfer e, however, the agent will deny his or her partner the aggregate payment P. This game is played every time the aggregate state is bad.

Notice that, for each agent, it is necessarily true that, in autarky, $P + e$ is preferred to P. Seen as a static one-shot game, this is a standard prisoner dilemma and {no, no} is its only Nash equilibrium.

We assume that the sequence of events is as follows. First, the state of the world gets revealed. Second, agents make their private transfers. Then, in bad aggregate states, they play the game just mentioned. Finally, they consume their

Table 6.3. *Second scheme: The payoff structure that rewards naysayers*

	2 says yes	2 says no
1 says yes	$\{P, P\}$	$\{0, P + e\}$
1 says no	$\{P + e, 0\}$	$\{e, e\}$

Table 6.4. *A noncredible scheme*

	2 says yes	2 says no
1 says yes	$\{P, P\}$	$\{-k, P + e\}$
1 says no	$\{P + e, -k\}$	$\{e - k, e - k\}$

disposable income (made of their endowment plus net private transfers, plus what they get from the government in case they play this game).

To analyze the welfare consequences of this game, we once again run some simulations. The first issue we have to solve is how to compute the equilibrium in a situation where the game played is not static but is played every time the bad aggregate state occurs. Under first-best allocation, it is trivial to show that, as long as e is small, playing {yes, yes} is the only equilibrium. We then compute the value of autarky by assuming that when the relationship breaks down, the two players play the game by taking into account what the other is likely to do in the future. It turns out that, in most situations, playing {no, no} is an equilibrium.[23] Under this assumption, the value of autarky is left, for small values of e, almost unaffected (and indeed slightly decreased because of the premium payed in good states). Therefore, we will have the same level of risk sharing achieved by our first scheme under the assumption that in autarky both agents would choose the "right" equilibrium, that is, {no, no}.

Such a scheme could be generalized to include a further punishment for the agent to whom the aggregate payment is denied, say k.

In this case the scheme would look like Table 6.4.

By making k large enough, one can decrease the value of autarky substantially. Therefore, the system gets closer and closer to the first-best allocation. However, even though we have not modeled it explicitly, it is not extremely realistic to assume that the World Bank or another centralized organization goes to a village affected by a bad aggregate shock and credibly threatens to remove resources from the agents in that village. This is why we prefer to stick to the simpler, if slightly less efficient, scheme we discuss herein.

6. CONCLUSIONS

Here we have discussed models of imperfect risk sharing. We have mainly focused on models where first-best allocation of resources might not be achieved because of the presence of enforceability problems. We have shown that these models provide a useful tool that allows us to characterize deviations from first best. Moreover, we have stressed that self-enforceable contracts are hybrids of

[23] Such an equilibrium is not likely to be unique. However, unicity problems are standard in repeated game frameworks. The game we are considering now, however, is much better than the previous one, in which we had multiple equilibria even when the game was considered to be a static one.

insurance and debt contracts and that the pattern of transfers they give rise to might differ substantially from those one would observe under first best.

We believe that this class of models is extremely useful and relevant to characterize poor, developing economies where information problems within the economy might be relatively few and yet few institutions that enforce contracts might exist. If, at the same time, it might be difficult to convey the information within the economy to agents to the outside (such as courts), it might be difficult to enter contracts that are not self-enforcing.

Given these considerations, any government intervention aimed at providing insurance within these village economies is bound to interfere with the working of private contracts and transfers. Indeed, we showed that there might be situations in which a well-meaning government might, by supplying insurance against aggregate shocks, so crowd out the private transfers as to make individuals worse off. In any case, it is important to understand the disruption in private markets that is generated by the introduction of government insurance schemes.

In the final section of the paper we tackled the issue of the optimal provision of aggregate insurance directly and showed that there might be large payoffs to the careful design of such a scheme. Not only can one design it in a way that avoids the crowding out, but, in doing so, one would actually improve the functioning of private markets. We constructed simple schemes, which could realistically be implemented by a fairly unsophisticated bureaucracy, that have this property.

Much work still has to be done, especially in empirically testing the implications of the models we have discussed.

ACKNOWLEDGMENTS

O. P. Attanasio's research was partly financed by a European Commission TMR grant on "New Approaches to the Study of Economic Fluctuations." J.-V. Ríos-Rull thanks the National Science Foundation for Grant SEC-0079504 and the University of Pennsylvania Research Foundation for their support. Thanks to Fabrizio Perri for his comments and for his help with the code. Tim Besley, Ethan Ligon, Narayana Kocherlakota, Tim Worrall, Mark Rosenzweig, Chris Udry, Tilman Borgers, Matthew Jackson, John Moore, Nobu Kiyotaki, and Robert Townsend provided useful comments. Jonathan Thomas was particularly helpful in discussing the optimal provision of aggregate insurance. Jose Gomez de Leon and Patricia Muniz were very helpful with the Progresa data used in this study. We also thank Pedro Albarran and Josep Pijoan-Mas for research assistance.

References

Abreu, D. (1988), "On the Theory of Infinite Repeated Games with Discounting," *Econometrica*, 56, 383–396.

Abreu, D., D. Pearce, and E. Stacchetti (1990), "Toward a Theory of Discounted Repeated Games with Imperfect Monitoring," *Econometrica*, 58, 1041–1063.

Aiyagari, S. R. (1994), "Uninsured Idiosyncratic Risk, and Aggregate Saving," *Quarterly Journal of Economics*, 109, 659–684.

Albarran, P. and O. Attanasio (2000), "Public Transfers and the Crowding Out of Private Transfers: Empirical Evidence from Mexico," mimeo, University College, London.

Alvarez, F. and U. Jermann (1998), "Asset Pricing When Risk Sharing Is Limited by Default," unpublished manuscript, University of Pennsylvania.

Atkeson, A. and R. E. Lucas (1992): "On efficient distribution with private information," *Review of Economic Studies*, 59, 427–453.

Atkeson, A. and M. Ogaki (1996), "Wealth Varying Intertemporal Elasticities of Substitution: Evidence from Panel and Aggregate Data," *Journal of Monetary Economics*, 38(3), 507–534.

Attanasio, O. and S. Davis (1996): "Relative Wage Movements and the Distribution of Consumption," *Journal of Political Economy*, 104, 1227–1262.

Attanasio, O. and J.-V. Ríos-Rull (2000a), "Consumption Smoothing in Island Economies: Can Public Insurance Reduce Welfare?," *European Economic Review*, 44, 1225–1258.

Attanasio, O. and J. V. Ríos-Rull (2000b), "On the Optimal Provision of Aggregate Insurance in the Presence of Enforceability Problems in the Provision of Private Insurance," mimeo, University College, London.

Attanasio, O. P., R. Blundell, and I. Preston (2000), "From Wage Shocks to Consumption Shocks," mimeo, University College, London.

Besley, T. (1995), "Property Rights and Investment Incentives: Theory and Evidence from Ghana," *Journal of Political Economy*, 103(5), 903–937.

Castaneda, A., J. Díaz-Giménez, and J.-V. Ríos-Rull (1998), "Exploring the Income Distribution Business Cycle Dynamics," *Journal of Monetary Economics*, 42(1), 93–130.

Castaneda, A., J. Díaz-Giménez, and J. V. Ríos-Rull (2002), "A General Equilibrium Analysis of Progressive Income Taxation: Quantifying the Trade-Offs," *Journal of Political Economy*, in press.

Coate, S. and M. Ravallion (1993), "Reciprocity Without Commitment: Characterization and Performance of Informal Insurance Arrangements," *Journal of Development Economics*, 40(1), 1–24.

Cochrane, J. H. (1991), "A Simple Test of Consumption Insurance," *Journal of Political Economy*, 99(5), 957–976.

Cole, H. L. and N. R. Kocherlakota (2001), "Efficient Allocations With Hidden Income and Hidden Storage," *Review of Economic Studies*, 68, 523–542.

Cox, D. (1987a), "Motives for Private Transfers," *Journal of Political Economy*, 95, 508–546.

Cox, G. (1987b), "Electoral Equilibrium Under Alternative Voting Institutions," *American Journal of Political Science*, 31, 82–108.

Cox, D., Z. Eser, and E. Jimenez (1998), "Motives for Private Transfers Over the Life Cycle: An Analytical Framework and Evidence for Peru," *Journal of Development Economics*, 55(1), 57–80.

Cox, D. and G. Jakubson (1995), "The Connection Between Public Transfers and Private Interfamily Transfers," *Journal of Public Economics*, 57, 129–67.

Dercon, S. and P. Krishnan (2000), "In Sickness and in Health: Risk Sharing Within Households in Rural Ethiopia," *Journal of Political Economy*, 108(4), 688–727.

Di Tella, R. and R. MacCulloch (1999), "Informal Family Assurance and the Design of the Welfare State," mimeo, Harvard University.

Fafchamps, M. (1999), "Risk Sharing and Quasi-Credit," *Journal of International Trade and Economic Development*, 8(3), 257–278.

Fafchamps, M. and S. Lund (2000), "Risk Sharing Networks in Rural Philippines," mimeo, Oxford University.

Foster, A. and M. Rosenzweig (1999), "Imperfect Commitment, Altruism, and the Family: Evidence from Transfer Behavior in Low-Income Rural Areas," mimeo, University of Pennsylvania.

Green, E. (1987), "Leading and the Smoothing of Uninsurable Income," in *Contractual Arrangements for Intertemporal Trade* (ed. by E. Prescott and N. Wallace), Minneapolis: University of Minnesota Press.

Hayashi, F., J. Altonji, and L. Kotlikoff (1996), "Risk-Sharing Between and Within Families," *Economica* 64(2), 261–294.

Heaton, J. and D. J. Lucas (1992), "Evaluating the Effects of Incomplete Markets on Risk Sharing and Asset Prices," Working Paper, Alfred P. Sloan School of Management, MIT.

Huggett, M. (1993), "The Risk Free Rate in Heterogeneous-Agents, Incomplete Insurance Economies," *Journal of Economic Dynamics and Control*, 17(5/6), 953–970.

Jackson, M. O. (2000), "Mechanism Theory," in *Encyclopedia of Life Support Systems*.

Jensen, R. (1999), "Public Transfers, Private Transfers, and the 'Crowding Out' Hypothesis: Evidence From South Africa," mimeo, Princeton University.

Judd, K. L. and J. Conklin (1993), "Computing Supergame Equilibria," mimeo, Stanford University.

Kehoe, P. and F. Perri (1997), "International Business Cycles with Endogenous Incomplete Markets," Working Paper, Federal Reserve Bank of Minnepolis.

Kehoe, T. and D. Levine (1993), "Debt Constrained Asset Markets," *Review of Economic Studies*, 60, 865–888.

Kehoe, T. J. and D. Levine (2001), "Liquidity Constrained vs. Debt Constrained Markets," *Econometrica*, 69, 575–598.

Kocherlakota, N. R. (1996), "Implications of Efficient Risk Sharing Without Commitment," *Review of Economic Studies*, 63(4), 595–609.

Krueger, D. and F. Perri (1999), "Risk Sharing: Private Insurance Markets or Redistributive Taxes?," Working Paper, Federal Reserve Bank of Minnepolis.

Krusell, P. and A. Smith (1997), "Income and Wealth Heterogeneity, Portfolio Choice, and Equilibrium Asset Returns," *Macroeconomic Dynamics*, 1(2), 387–422.

Krusell, P. and A. Smith (1998), "Income and Wealth Heterogeneity in the Macroeconomy," *Journal of Political Economy*, 106, 867–896.

Lambertini, L. (1999), "Borrowing and Lending Without Commitment and with Finite Life," mimeo, UCLA.

Ligon, E., J. P. Thomas, and T. Worrall (2000), "Mutual Insurance, Individual Savings and Limited Commitment," *Review of Economic Dynamics*, 3(3), 1–47.

Ligon, E., J. P. Thomas, and T. Worrall (2001), "Informal Insurance Arrangements in Village Economies," *Review of Economic Studies*, in press.

Luttmer, E. (1999), "What Level of Fixed Costs Can Reconcile Consumption and Stock Returns?," *Journal of Political Economy*, 107(5), 969–997.

Mace, B. J. (1991), "Full Insurance in the Presence of Aggregate Uncertainty," *Journal of Political Economy*, 99(5), 928–956.

Marcet, A. and R. Marimon (1992), "Communication, Commitment and Growth," *Journal of Economic Theory*, 58(2), 219–249.

Marcet, A. and R. Marimon (1995), "Recursive Contracts," unpublished manuscript, Universitat Pompeu Fabra.

Marcet, A. and K. J. Singleton (1990), "Equilibrium Assets Prices and Savings of Heterogeneous Agents in the Presence of Portfolio Constraints," mimeo, Carnegie Mellon University.

Moore, J. (1992), "Implementation, Contracts, and Renegotiation in Environments With Complete Information," in *Advances in Economic Theory: Sixth World Congress*, Vol. 1 (ed. by J.-J. Laffont), *Econometric Society Monographs*, Cambridge: Cambridge University Press, 182–192.

Moore, J. and R. Repullo (1988), "Subgame Perfect Implementation," *Econometrica*, 58, 1083–1099.

Moulin, H. (1986), *Game Theory for the Social Sciences*. New York: New York University Press.

Ogaki, M. and Q. Zhang (2001), "Decreasing Risk Aversion and Tests of Risk Sharing," *Econometrica*, 69, 515–26.

Phelan, C. and E. Stacchetti (2001), "Sequential Equilibria in a Ramsey Tax Model," *Econometrica*, 69(6), 1491–1518.

Phelan, C. and R. M. Townsend (1991), "Computing Multiperiod Information Constrained Optima," *Review of Economic Studies*, 58, 853–881.

Platteau, J. P. (1997), "Mutual Insurance as an Elusive Concept in Traditional Rural Communities," *Journal of Development Studies*, 33, 764–796.

Platteau, J. P. and A. Abraham (1987), "An Inquiry Into Quasi-Credit Contracts: The Role of Reciprocal Credit and Interlinked Deals in Small-Scale Fishing Communities," *Journal of Development Studies*, 23, 461–490.

Ravallion, M. and S. Chaudhuri (1997), "*Risk and Insurance in Village India:* Comment," *Econometrica*, 65(1), 171–184.

Telmer, C. I. (1992), "Asset Pricing Puzzles and Incomplete Markets," Working Paper, Queen's University.

Thomas, J. and T. Worrall (1988), "Self-Enforcing Wage Contracts," *Review of Economic Studies*, 55, 541–554.

Thomas, J. and T. Worrall (1990), "Income Fluctuations and Asymmetric Information: An Example of a Repeated Principal-Agent Problem," *Journal of Economic Theory*, 51(2), 367–390.

Townsend, R. M. (1994), "Risk and Insurance in Village India," *Econometrica*, 62(3), 539–591.

Udry, C. (1994), "Risk and Insurance in a Rural Credit Market: An Empirical Investigation in Northern Nigeria," *Review of Economic Studies*, 61(3), 495–526.

Wang, C. and S. Williamson (1996), "Unemployment Insurance With Moral Hazard in a Dynamic Economy," *Carnegie-Rochester Conference Series on Public Policy*, 44, 1–41.

CHAPTER 7

Computational Methods for Dynamic Equilibria with Heterogeneous Agents

Kenneth L. Judd, Felix Kubler, and Karl Schmedders

1. INTRODUCTION

Computational methods have become increasingly important in the analysis of dynamic general equilibrium problems. These methods are being used, for example, to study the incidence of tax and monetary policies in dynamic models of growth, commodity storage in various models of agricultural commodity markets, and price formation in dynamic models of asset markets. Many early computational methods relied primarily on intuitive economic tatonnement stories, and produced moderately successful algorithms. Even when these methods worked, they were usually slow. Furthermore, as we know from general equilibrium theory, tatonnement methods may not converge even with good initial guesses. In the past decade, the computational literature has made more use of formal mathematical tools from numerical analysis and perturbation theory. This use has resulted in more powerful algorithms that can attack increasingly complex problems. These developments are particularly important when we try to solve models with several agents. This essay reviews the key ideas used in recent work, gives some examples of their advantages, and indicates the likely directions future work will take.

It is particularly appropriate that the 2000 World Congress of the Econometric Society include a survey of recent computational literature, because computational methodology is inherently an important part of what is broadly called "econometrics." Ragnar Frisch, in his editorial in the initial issue of *Econometrica*, defined econometrics as the "unification of the theoretical-quantitative and the empirical-quantitative approach to economic problems." He said:

> This emphasis on the quantitative aspect of economic problems has a profound significance. Economic life is a complex network of relationships operating in all directions. Therefore, so long as we confine ourselves to statements in general terms about one economic factor having an effect on some other factor, almost any sort of relationship may be selected, postulated as a law, and explained by a plausible argument. Thus, there exists a real danger of advancing statements and conclusions, which – although true as tendencies in a very restricted sense – are nevertheless thoroughly inadequate, or even

> misleading if offered as an explanation of the situation. To use an extreme illustration, they may be just as deceptive as to say that when a man tries to row a boat forward, the boat will be driven backward because of the pressure exerted by his feet. The rowboat situation is not, of course, explained by finding out that there exists a pressure in one direction or another, but only by comparing the relative magnitudes of a number of pressures and counter-pressures. It is this comparison of magnitudes that gives a real significance to the analysis. Many, if not most, of the situations we have to face in economics are of just this sort.

Dynamic general equilibrium problems are excellent examples of problems with "a complex network of relationships operating in all directions." Recent work on computational methods for dynamic models shows the value of uniting economic theory and mathematics to create a quantitative analysis of economic problems, and this essay reviews some of these developments. Frisch also said that

> [Mathematics] is indispensable in a great many cases. Many of the essential things in the new setting of the problems are so complex that it is impossible to discuss them safely and consistently without the use of mathematics.

Frisch's emphasis on the necessity of mathematics is relevant here because there are examples (we will give one, but it is not unique) where intuitive ad hoc schemes lacking a proper mathematical foundation can give unreliable answers to economic questions. Fortunately, there are mathematically sound methods from numerical analysis that can be used instead.

In this paper we address the problem of computing equilibria of dynamic economic models, with special attention to the problems that arise when there are several agents. Some of these techniques are straightforward generalizations of methods applied to representative agent models, but heterogeneous agent models often present new problems requiring new techniques. We stress that this survey focuses on recent advances in computational methods; we do not attempt to survey all applications in the applied dynamic general equilibrium literature.[1] Dynamic models with heterogeneous agents are inherently difficult to exposit precisely and compactly. Therefore, we use simpler models to illustrate many computational concepts and then indicate how they have been applied more generally to heterogeneous-agent models.

Most dynamic problems in economics have a structure that numerical methods can exploit. Some problems are time homogeneous and some have time dependencies; we will distinguish between the two cases because computational methods differ. Our discussion is organized as follows. Section 2 discusses recent developments in solving perfect foresight models. They are extensively used in the applied macroeconomics and dynamic computational general equilibrium literatures to examine problems with many state variables or in which

[1] This paper focuses on computational ideas that have been developed in recent years. Therefore, we focus on papers that present and describe in some detail their computational ideas and methods.

calendar time enters into the analysis because of, for example, partially anticipated changes in policy or environment.

Many economic problems, such as real business cycle models, have a time-homogeneous character and have a small number of state variables. Section 3 presents simple examples of time-homogeneous models of moderate size that we use in our exposition. The projection method from the numerical analysis literature gives us a useful framework of general concepts within which we can discuss and compare most methods used by economists. The projection method also suggests many new, potentially more powerful, algorithms. Section 4 presents the details of projection methods for solving functional equations. Section 5 provides some details on how to apply projection methods for stationary dynamic economic models. Section 6 describes an infinite-horizon model with finitely many agents and incomplete markets, and is an example of asset-pricing problems that have received much attention in the past decade. It is difficult to compute equilibria in these models, and the various methods used in this literature illustrate the evolution of computational methodology in economics. We use this asset model as a way of illustrating the key computational difficulties present in many dynamic economic models and how we can address these problems. Section 7 presents methods for solving the dynamic incomplete asset market model that combine continuous approximations of pricing and trading strategies, new methods from the computable general equilibrium literature on solving models with incomplete asset markets, and projection methods. Section 8 presents an example illustrating the computational difficulties that arise naturally in asset market models. Section 9 describes earlier methods developed for similar dynamic asset-pricing models and compares them with the more recent approaches.

Recent years have also seen progress in numerical methods for solving dynamic games. Section 10 presents some recent solution methods for problems in which strategic concerns are important. This includes problems in which agents, such as oligopolists, have market power, and problems in which actors, such as governments, face dynamic consistency problems and cannot precommit to future actions.

Perturbation methods give us an alternative approach to solving dynamic models. The most familiar example of this approach is the linearization methods described by Magill (1977) and Kydland and Prescott (1982). Recent work has adapted methods from the mathematics literature to show how to compute more general and robust Taylor series expansions of equilibrium relations. Section 11 presents perturbation and asymptotic methods for solving dynamic economic models. Perturbation methods are particularly valuable for analyzing models with too many state variables for projection methods to handle and where we know that equilibrium will not wander too much from some central states.

2. PERFECT FORESIGHT MODELS

Perfect foresight models are often used to analyze dynamic economic questions, and they were the first models for which numerical methods were developed.

The typical model has a relatively simple dynamic structure. Let $x_t \in \mathbb{R}^n$ be a list of time t values for economic variables such as consumption, labor supply, capital stock, output, wages, and so on, and let $z_t \in \mathbb{R}^m$ be a list of exogenous variables, such as productivity levels, tax rates, and monetary growth rates, at time t. Perfect foresight models have the form

$$g(t, X, Z) = 0, \quad t = 0, 1, 2, \ldots, \tag{2.1}$$

$$x_{0,i} = \bar{x}_{0,i}, \quad i = 1, 2, \ldots, n_I, \tag{2.2}$$

$$x_t \text{ bounded}, \tag{2.3}$$

where

$$X \equiv (x_0, x_1, x_2, \ldots, x_s, \ldots),$$
$$Z \equiv (z_0, z_1, z_2, \ldots, z_s, \ldots),$$

and $g(t, X, Z) : \mathbb{R} \times \mathbb{R}^{n\times\infty} \times \mathbb{R}^{m\times\infty} \to \mathbb{R}^n$ is a collection of n functions representing equilibrium. The equations in (2.1) include Euler equations, market-clearing conditions, and any other equations in the definition of equilibrium. Some economic variables have fixed predetermined values at $t = 0$ represented by the $n_I < n$ conditions in (2.2). Boundedness conditions in (2.3) provide additional conditions that tie down equilibrium. We need to find a bounded sequence of values for x_t satisfying all the equations in (2.1) and (2.2).

A simple example is the optimal growth problem

$$\max_{c_t} \sum_{t=0}^{\infty} \beta^t u(c_t) \tag{2.4}$$
$$\text{s.t. } k_{t+1} = F(k_t) - c_t,$$
$$k_0 = \bar{k}_0.$$

The solution to (2.4) satisfies the Euler equation $u'(c_t) = \beta u'(c_{t+1})F'(k_{t+1})$. In the notation of (2.1) and (2.2), we define $x \equiv (c, k)$ (there are no exogenous variables) and express the solution to (2.4) as

$$g_1(t, X) \equiv u'(c_t) - \beta u'(c_{t+1})F'(k_{t+1}) = 0, \quad t = 0, 1, 2, \ldots, \tag{2.5}$$
$$g_2(t, X) \equiv k_{t+1} - F(k_t) + c_t = 0, \quad t = 1, 2, \ldots,$$
$$k_0 = \bar{k}_0.$$

The capital stock has a predetermined value at $t = 0$. We shall use (2.5) as an example. Equation (2.5) is a problem with one type of agent and one good, but the approach can be used to analyze more general models. When we have heterogeneous agents, multiple goods, multiple sectors, and/or multiple countries, equilibrium consists of Euler equations for each type of agent for each decision variable, market-clearing conditions for each market, and any other equilibrium conditions. These are all stacked into the list $g(t, X, Z)$ in (2.1). Perfect foresight models are used to examine stochastic problems by allowing

the z_t to represent shocks and then solving (2.1) and (2.2) for many possible realizations[2] of Z.

This example is far simpler than the models used by policy analysts, such as MULTIMOD and models developed at the U.S. Federal Reserve Bank, which have many goods, many groups of agents, many countries, monetary policy shocks, and/or fiscal policy. Sometimes these models ask questions such as, What happens if we change policy in five years?, which generates a problem that depends on calendar time. These problems often have several predetermined state variables and several free variables such as consumption, labor supply, and the price level. The first large, rational expectations macroeconometric models were perfect foresight models of the form in (2.1)–(2.3). The Fair–Taylor (1983) method was the first one developed to analyze these large models. More recently, economists have applied methods from the mathematical literature on solving large systems of equations, and they have applied projection methods to perfect foresight models. This section reviews and compares some of the methods proposed for solving perfect foresight models.

2.1. General Considerations

Perfect foresight models are essentially nonlinear equations in $\mathbb{R}^\infty$. The forward-looking aspect of dynamic general equilibrium analysis creates links between current and future economic variables, and it generates an infinite system of nonlinear equations with an infinite number of unknowns. Under some conditions, there will be a locally unique solution. For example, (2.5) has a unique solution for any k_0. All methods we discuss assume local uniqueness.

These models are often Arrow–Debreu general equilibrium problems, but their large size makes conventional computational general equilibrium procedures such as Scarf's algorithm and homotopy procedures impractical. Any solution method must reduce the problem in some way. Most methods use *time truncation* to reduce the problem to a finite-horizon problem. That is, they solve the truncated problem

$$g(t, x_0, x_1, \ldots, x_T, x^*, x^*, \ldots, Z) = 0, \quad t = 0, 1, 2, \ldots, T, \tag{2.6}$$

$$x_{0,i} = \bar{x}_{0,i}, \quad i = 1, 2, \ldots, n_I, \tag{2.7}$$

where x^* is the steady-state value of x. Some components of x_T are also fixed at their long-run values to make the number of equations in (2.6) and (2.7) equal to the number of unknowns. Time truncation reduces (2.1)–(2.3) to a system of nT nonlinear equations in nT unknowns. There is no boundedness equation in (2.6) and (2.7) because (2.6) imposes $x_t = x^*$ for $t > T$. Because T must be large in order to be an acceptable approximation for the total dynamic process, we still cannot use conventional methods.

[2] See Fair and Taylor (1983) for an example of this approach to solving stochastic rational expectations models.

There is always the question of what T should be. Any method should try alternative values for T and accept a solution only when the choice of T does not substantially affect the solution, a step that can add substantial computational effort.

2.2. Gauss–Jacobi and Gauss–Seidel Methods

Perfect foresight models of the form

$$g(t, x_t, x_{t+1}, Z) = 0 \tag{2.8}$$

are solved by using methods from the literature on solving large systems of equations. Fair and Taylor (1983) introduced an intuitive approach. They begin with an initial guess $X^0 = (x_1, x_2, \ldots, x_T, x^*, x^*, \ldots)$, which incorporates the time-truncation approach. Then they use the time t equation $g(t, x_t, x_{t+1}, Z) = 0$ to compute a new guess for x_t given the initial guess for x_{t+1}^0. In general, the $(i+1)$st guess for X, denoted X^{i+1}, is constructed componentwise by solving

$$g\big(t, x_t^{i+1}, x_{t+1}^i, Z\big) = 0, \quad t = 1, 2, \ldots \tag{2.9}$$

for x_t^{i+1}, the time t component of X^{i+1}. Their scheme is a block Gauss–Jacobi scheme, because only elements of X^i are used to compute X^{i+1}. Solving for x_t^{i+1} given x_{t+1}^i in (2.9) is also a nonlinear equation, but it is generally of moderate size and solvable by conventional schemes, such as those by Newton or Gauss–Seidel. They also suggest that one try different truncation times T until changes in T create small changes in the solution. The Fair–Taylor scheme is reliable but tends to be slow because of the Gauss–Jacobi structure. The slow speed makes it difficult to solve with high accuracy because tight accuracy targets would require too many iterations. Also, Gauss–Jacobi schemes may not converge even if one begins with a good initial guess.

Because T is typically large, we need to develop special methods. Fortunately, we can apply methods from the literature on solving large systems (see, e.g., Kelley, 1995, Saad, 1996, and Young, 1971). Some schemes reorder the equations in order to accelerate convergence. Some examples of this approach are given by Fisher and Hughes Hallett (1987, 1988) and Hughes Hallett and Piscitelli (1998). Convergence of such methods depends on the ordering of the equations and is linear at best. The advantages are their simplicity and small memory requirements. However, they may not converge even after the equations have been reordered and several dampening factors have been tried.

2.3. Newton-Style Methods

More recently, some authors have used Newtonian and related methods to solve dynamic economic problems. Newton's method for solving the system of equations $g(x) = 0$ is the iteration $x^{k+1} = x^k - J(x^k)^{-1} g(x^k)$. Newton's method converges rapidly if the initial guess x^0 is good. Unfortunately, Newton's method is impractical for general large systems because the Jacobian of a

system of n equations has n^2 derivatives, an impractically large amount of computation if n is large. However, Newton methods can be applied to models with the simple lag structure in (2.8) because the Jacobian for perfect foresight problems is sparse; that is, most elements are zero.

The LBJ algorithm (see Laffargue, 1990, Boucekkine, 1995, Juillard, 1996, and Juillard et al., 1998) takes notice of a special structure in many perfect foresight models and exploits it to apply Newton's method. Because the time t equation $g(t, x_t, x_{t+1}, Z) = 0$ involves only x_t and x_{t+1}, each row in a Jacobian involves only a small fraction of all the unknowns. Let $g_i(t, x_t, x_{t+1})$ denote $\partial g(t, x_t, x_{t+1})/\partial x_i$. The Jacobian of (2.6) and (2.7) is

$$J(x) = \begin{bmatrix} g_1(1, x_1, x_2) & g_2(1, x_1, x_2) & 0 & \dots \\ 0 & g_2(2, x_2, x_3) & g_3(2, x_2, x_3) & \dots \\ 0 & 0 & g_3(3, x_3, x_4) & \dots \\ 0 & 0 & 0 & \dots \\ 0 & 0 & 0 & \ddots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix}$$

and is a sparse matrix for large n. Specifically, $J(x)$ is nearly diagonal here in the sense that all nonzero elements of $J(x)$ are within n columns of the diagonal even though there are nT columns in $J(x)$.

This fact can be used in a Newtonian approach. More precisely, iteration $k + 1$ of Newton's method solves

$$\begin{aligned} J(x^k)\Delta &= -g(x^k), \\ x^{k+1} &= x^k + \Delta. \end{aligned} \tag{2.10}$$

Because $J(x)$ is sparse, one can use sparse matrix methods to solve the linear equation $J(x^k)\Delta = -g(x^k)$ for the Newton step Δ. Juillard et al. (1998) examined a Newtonian strategy exploiting this sparseness and were able to solve large problems faster than they could by using Gaussian methods, such as the Fair–Taylor method, and do so with high accuracy.

Solving (2.10) can be difficult if x and T are large even if the Jacobian $J(x)$ is sparse. Gilli and Pauletto (1998) economize on this by using Krylov methods to compute an approximate solution to (2.10). An approximate solution is adequate because the important thing is to arrive at some Δ that takes the iteration in the right direction. Krylov methods find an approximate solution by projecting (2.10) into a smaller dimension and solving the projected problem. Gilli and Pauletto (1998) report significant gains in algorithm speed over sparse Newtonian methods.

Some Gaussian methods have an economic motivation, often turning on learning ideas. For example, one way to interpret Fair–Taylor is to say that agents compute their actions given expectations, but then those computed actions form the next set of expectations. Although the storytelling approach to solving dynamic economic models has some intuitive appeal, it produces algorithms that

converge linearly if at all. Although Newton's method and other methods from the numerical analysis literature have no obvious economic "story," they bring the possibility of more rapid and/or reliable convergence and more accurate solutions.

2.4. Parametric Path Method

The parametric path approach proposed in Judd (2002) employs a substantially different strategy to solve (2.1) and (2.2). Instead of treating each component of $X = (x_0, x_1, \ldots)$ as independent, it uses information about how the true value of x_t evolves over time. For example, the sequence 1, 2, 1, 2, ... is not likely to represent a quarterly series for the capital stock or even aggregate consumption. Capital stock sequences will be relatively smooth because the capital stock cannot change quickly. Consumption sequences are also likely to be smooth when agents have concave utility functions. This feature of the solutions is not exploited by standard methods because they treat each distinct x_t separately. Instead, our intuition says that the sequence $(x_0, x_1, \ldots)$ should be a smooth function of time t. This insight allows us to reduce (2.1) and (2.2) to a much smaller system to which we can apply methods not practical for these equations.

The key idea behind the parametric path method can be illustrated in its application to (2.5). Theory tells us that the equilibrium capital sequence for (2.5) converges to the steady state at linear rate λ, where λ is the stable eigenvalue of the linearization of (2.5) around the steady state k^{ss}. We also know that the time path of capital is "smooth" and that convergence is asymptotically monotone. This, together with the initial condition $k(0) = k_0$, suggests the parameterization

$$K(t;a) = \left(k_0 + \sum_{j=1}^{m} a_j t^j\right) e^{-\lambda t} + k^{ss}(1 - e^{-\lambda t}). \tag{2.11}$$

There are two key features of (2.11). First, for any a, $K(t;a)$ converges to k^{ss} because of the exponential decay term $e^{-\lambda t}$. Second, $k_0 = K(0;a)$ for any a. Therefore, (2.11) automatically satisfies both the initial conditions and the boundedness condition for any a. These facts allow us to focus on finding an $a \in \mathbb{R}^m$ that produces a good approximate solution to (2.5) without getting sidetracked by convergence problems.

System (2.5) is equivalent to the second-order difference equation

$$u'(F(k(t)) - k(t+1)) = \beta u'(F(k(t+1)) - k(t+2))F'(k(t+1)).$$

We want to approximate $k(t)$ with $K(t;a)$ for some a. Therefore, the parametric

path method defines the residual function

$$R(t;a)=u'(F(K(t;a)) - K(t+1;a)) - \beta u'(F(K(t+1;a)) - K(t+2;a))F'(K(t+1;a))$$

and searches for an $a \in \mathbb{R}^m$ that makes $R(t;a)$ close to being zero for all t. Note that $R(t;a)$ is well defined for any real value of t, not just the integers, because $k(t)$ is defined for all t in (2.11). Because $K(t;a) \to k^{ss}$ as $t \to \infty$, $R(t;a) \to 0$ as $t \to \infty$. Therefore, the Euler equation is satisfied asymptotically for any a, allowing us to focus on making $R(t;a)$ small at finite values of t. To identify the coefficients a, we define the set of projection formulas

$$P_j(a) = \sum_{t=0}^{\infty} R(t;a)t^i e^{-\lambda t} dt, \quad i = 0, 1, \ldots. \tag{2.12}$$

The summation in (2.12) is infinite, but by combining orthogonal polynomial theory and appropriate changes of variables, Judd (2002) derives good choices of weights ω_i and times t_i such that

$$\widehat{P}_j(a) = \sum_{i=0}^{N} \omega_i R(t_i;a)t_i^j, \quad j = 0, 1, \ldots$$

is a good approximation of (2.12) for some weights ω_i and times, and then uses Newton's method to find coefficients $a \in \mathbb{R}^m$ that solve the system

$$\widehat{P}_j(a) = 0, \quad j = 1, \ldots, m.$$

This is a simple example of the projection method described in greater detail in the sections that follow.

There has been steady progress in solving large perfect foresight systems, and we expect progress to continue. The new developments have a common approach. They all exploit the dynamic structure of the problem more extensively than did the Fair–Taylor procedure, and they also bring appropriate methods from the numerical analysis literature on solving large problems and approximating unknown functions.

3. TIME-HOMOGENEOUS DYNAMIC ECONOMIC PROBLEMS

The kinds of problems discussed in Section 2 can be used to analyze large economic systems, but they are limited to problems that are perfect foresight models or close to being perfect foresight models. One can, as outlined in Fair and Taylor, extend the analysis to stochastic contexts, but this leads to solving several perfect foresight problems, each with different realizations of random shocks. This approach to stochastic modeling is somewhat limiting and costly to execute. The next set of methods aims at solving fully stochastic models in an efficient fashion.

The key fact we will exploit now is that many dynamic economic models take a stationary form; that is, calendar time does not affect the equilibrium. Equilibrium of stationary problems can be expressed in feedback rules, expressing the free endogenous variables, such as prices, consumption, and labor supply, as functions of the predetermined variables, such as capital stocks and lagged productivity levels. They also often involve uncertainty about productivity, policy, or other exogenous economic factors. These models take the form

$$\begin{aligned} 0 &= E\{g(x_t, y_t, x_{t+1}, y_{t+1}, z_{t+1}) \mid x_t\}, \\ x_{t+1} &= F(x_t, y_t, z_t), \end{aligned}$$

where x_t is a vector of variables that are predetermined at the beginning of period t, y_t are the free variables, and z_t are shocks to the system. The function $F(x, y, z)$ is the law of motion for the predetermined variables, and g is a list of equilibrium conditions such as Euler equations.

When a model is time homogeneous, equilibrium is characterized by some equilibrium rule, $y_t = Y(x_t)$, which expresses the value of the free variables in terms of the state x such that

$$E\{g(x, Y(x), F(x, Y(x), z), Y(F(x, Y(x), z)) \mid x\} = 0 \tag{3.1}$$

holds for all values of x. The equilibrium rule $Y(x)$ expresses variables such as consumption and prices as functions of the state x. This focus contrasts with the approach of the previous section, where we aimed at solving the stochastic sequence of equilibrium prices, consumption, and so on. Both problems, finding the sequence x_t and finding the decision rule $Y(x)$, are infinite dimensional, but they lie in different spaces and the solutions use different tools. This approach is clearly impractical if the state variable x is large, but this approach is better for modeling stochastic economies.

For example, the stochastic version of (2.4), investigated in the Taylor–Uhlig (1990) symposium and in Judd (1992), is

$$\begin{aligned} &\max_c E\left\{\sum_{t=0}^{\infty} \beta^t u(c_t)\right\}, \\ &k_{t+1} = F(k_t, \theta_t) - c_t, \\ &\ln \theta_{t+1} = \rho \ln \theta_t + \epsilon_{t+1}, \end{aligned} \tag{3.2}$$

where k_t is the beginning-of-period capital stock, θ_t is a productivity parameter with iid productivity shocks $\epsilon_t \sim N(0, \sigma^2)$, and $F(k, \theta)$ is the gross production function. In this problem, both k and θ are needed for a sufficient description of the state. Hence, consumption is a function, $C(k, \theta)$, of both k and θ, and the Euler equation is

$$\begin{aligned} u'(C(k,\theta)) = \beta E\{&u'(C(F(k,\theta) - C(k,\theta), \tilde{\theta}))F_k(F(k,\theta) \\ &- C(k,\theta), \tilde{\theta})|\theta\}. \end{aligned} \tag{3.3}$$

The case of two agents in a competitive economy can be similarly analyzed. Suppose that type i agents have utility

$$E\left\{\sum_{t=0}^{\infty}\beta^t u_i(c_{i,t})\right\}$$

and budget constraint $k_{i,t+1} = R_t k_{i,t} + w_t - c_{i,t}$, where k_i is the amount of capital stock owned by the representative agents of type i, R_t is the random return from capital, and w is the wage from the supply of one unit of labor. Here the state variable is the capital stock owned by each type as well as the productivity level. In this case, equilibrium consumption of type i agents is a function of the distribution of wealth; let $C_i(k_1, k_2, \theta)$ be the consumption of type i agents when the wealth distribution is $k = (k_1, k_2)$ and the productivity level is θ. We assume that equity is the only asset that can be held; the more general case is examined in a later section. The equilibrium is defined by the collection of Euler equations

$$\begin{aligned}
&u_i'(C^i(k_1^+, k_2^+, \theta)) = \beta E\{u'(C^i(k_1^+, k_2^+, \tilde{\theta}))R(k^+, \tilde{\theta})|\theta\}, \quad i = 1, 2,\\
&k_i^+ = Y^i(k, \theta) - C^i(k, \theta), \quad i = 1, 2,\\
&Y^i(k, \theta) = k_i R(k, \theta) + w(k, \theta), \quad i = 1, 2,\\
&R(k, \theta) = F'(k_1 + k_2),\\
&w(k, \theta) = F(k_1 + k_2) - (k_1 + k_2)F'(k_1 + k_2),
\end{aligned} \tag{3.4}$$

where $Y(k, \theta) \in \mathbb{R}^2$ is the distribution of income in a period with initial capital stock distribution k and productivity θ, and wages are $w(k, \theta)$. This example is a simple one, but it illustrates the basic features of models with heterogeneous agents. In particular, consumption and other decisions depend on the distribution of income across agents. This example has only one asset. More generally, we would like to examine the case in which there are multiple assets. In that case, the state variable is even larger because the distribution of holdings of each asset may be important.

This paper discusses various numerical methods for solving such models proposed in recent years. The paper discusses both perturbation and projection methods.

4. GENERAL PROJECTION ALGORITHM

Solving functional equations of the sort in (3.4) initially appears to be difficult because it is still an infinite-dimensional problem. Numerical rational expectations methods, beginning with that by Gustafson (1958), focus on finite-dimensional approximations of policy functions and other important functions, and then implement some sort of iterative procedure to find a finite-dimensional approximation that nearly solves the functional equations defining equilibrium.

For example, solutions to (3.2) typically use approximations of the form

$$\widehat{C}(k,\theta) = \sum_{i=0}^{n} a_i \varphi_i(k,\theta),$$

where the φ_i comprise a basis for the space of functions thought to contain the solution, and focus on finding good choices for the a coefficients. This is true of both perturbation and projection methods. The next section describes the ideas behind projection methods. We discuss perturbation methods in a later section.

Most methods used to solve dynamic economic models are examples of what are called projection methods in the mathematics literature.[3] Suppose that economic analysis shows that equilibrium can be expressed as an operator equation

$$\mathcal{N}(f) = 0,$$

where f is a function $f : D \subset \mathbb{R}^N \to \mathbb{R}^M$, $\mathcal{N}$ is an operator $\mathcal{N} : B_1 \to B_2$, and the B_i are function spaces. The unknown function f expresses prices, consumption decisions, and similar economic quantities as a function of state variable x. Typically, $\mathcal{N}$ expresses equilibrium conditions such as Euler equations and consists of a composition of algebraic operations, differential and integral operators, and functional compositions, and is frequently nonlinear. We show how to implement the canonical projection technique in a step-by-step fashion.

The first step decides how to represent approximate solutions. Here we assume that we build the approximation $\hat{f}$ from linear combinations of simple functions, but nonlinear representations are also possible. We also need a concept of when two functions are close. Therefore, the first step is to choose a basis and an appropriate concept of distance.

1. Step 1: Choose bases, $\Phi_j = \{\varphi_i\}_{i=1}^{\infty}$, and inner products, $\langle \cdot, \cdot \rangle_j$, over B_j, $j = 1, 2$.

There are many criteria that the basis and inner product should satisfy. The basis Φ_1 should be "rich." In particular, it should be complete in B_1. We generally use inner products of the form

$$\langle f(x), g(x) \rangle_1 \equiv \int_D f(x) g(x) w(x) dx$$

for some weighting function $w(x) \geq 0$. Computational considerations say that the φ_i should be simple to compute, and similar in size to avoid scaling problems. The basis elements should "look something like" the solution. In particular, we should use smooth functions to approximate smooth functions, but use splines to approximate functions that may have kinks or other extreme local behavior.

[3] The term "projection method" is a catchall term in the mathematical literature, which includes several methods including the method of weighted residuals, finite-element methods, Galerkin methods, the least-squares method, and the Rayleigh–Ritz method.

Because of its special properties, a generally useful choice is the Chebyshev polynomial family. If, in contrast, one has a basis that is known to efficiently approximate the solution, one should use that instead or combine it with a standard orthogonal family.

Next, we decide how many basis elements to use and how to implement $\mathcal{N}$.

2. Step 2: Choose a degree of approximation n, a computable approximation $\hat{\mathcal{N}}$ of $\mathcal{N}$, and a collection of n test functions from B_2, $p_i : D \to \mathbb{R}^M$, $i = 1, \ldots, n$. Define $\hat{f}(x) \equiv \sum_{i=1}^{n} a_i \varphi_i(x)$.

The best choice of n cannot be determined a priori. One initially begins with small n and increases n until some diagnostic indicates that little is gained by continuing. Similar issues arise in choosing $\hat{\mathcal{N}}$. Sometimes we can take $\hat{\mathcal{N}} = \mathcal{N}$, but more generally some approximation is necessary. The test functions p_i are used to identify the unknown coefficients a.

Step 1 lays down the topological structure and Step 2 fixes the flexibility of the approximation. Once we have made these basic decisions, we begin our search for an approximate solution to the problem. Because the true solution f satisfies $\mathcal{N}(f) = 0$, we search for some $\hat{f}$ that makes $\hat{\mathcal{N}}(\hat{f})$ "nearly" equal to the zero function. Because $\hat{f}$ is parameterized by a, the problem reduces to finding a coefficient vector a that makes $\hat{\mathcal{N}}(\hat{f})$ nearly zero. This search for a is the focus of Steps 3–5.

3. Step 3: For a guess a, compute the approximation, $\hat{f} \equiv \sum_{i=1}^{n} a_i \varphi_i(x)$, and the residual function,

$$R(x; a) \equiv (\hat{\mathcal{N}}(\hat{f}))(x).$$

The first guess of a should reflect some initial knowledge about the solution. After the initial guess, further guesses are generated in Steps 4 and 5, where we see how we use the inner product, $\langle \cdot, \cdot \rangle_2$, defined in the space B_2, to define what "near" means.

4. Step 4: For each guess of a, compute the n projections,

$$P_i(a) \equiv \langle R(\cdot; a), p_i(\cdot) \rangle_2, \quad i = 1, \ldots, n,$$

or the L^2 norm $\langle R(x; a), R(x; a) \rangle$.

Step 4 reduces the original infinite-dimensional problem to a finite-dimensional problem. Step 5 finishes the job.

5. Step 5: By making a series of guesses over a and iterating over Steps 3 and 4, find a value for a that sets the n projections equal to zero or minimizes the L^2 norm of $R(x; a)$.

There are many ways to implement the ideas in Steps 3–5. First, the *least-squares* approach chooses a to minimize the "weighted sum of squared residuals":

$$\min_a \langle R(x; a), R(x; a) \rangle.$$

Least-squares methods are easy to implement and can use optimization software directly. Unfortunately, they often perform poorly because there may be local minima that are not global minima, and the objective may be poorly conditioned.

Although the least-squares method is a direct approach to making $R(x;a)$ small, most projection techniques find approximations by fixing n projections and choosing a to make the projection of the residual function in each of those n directions zero. Formally, these methods find a such that $\langle R(x;a), p_i(x)\rangle_2 = 0$ for some specified collection of functions, p_i. This reduces the problem to a set of nonlinear equations. Different choices of the p_i define different implementations of the projection method.

The *Galerkin method* is one such method and uses the first n elements of the basis for the projection directions. The coefficient vector a is a solution to the following equations:

$$P_i(a) \equiv \langle R(x;a), \varphi_i(x)\rangle = 0, \quad i = 1, \ldots, n.$$

A *collocation method* takes n points from the domain D, $\{x_i\}_{i=1}^n$, and it chooses a to solve

$$R(x_i;a) = 0, \quad i = 1, \ldots, n.$$

Orthogonal collocation sets the x_i to be the zeros of the nth basis element, where the basis elements are orthogonal with respect to the inner product. The performance of Chebyshev collocation is often surprisingly good; see, for example, Judd (1992).

Choosing the projection conditions is a critical decision because the major computational task is the computation of those projections. The collocation method is fastest in this regard because it uses only the value of R at n points. More generally, we generally require numerical quadrature techniques to compute the inner products in $P(a)$. A typical quadrature formula approximates

$$\int_a^b f(x)g(x)dx$$

with a finite sum

$$\sum_{i=1}^n \omega_i f(x_i),$$

where the x_i are the quadrature nodes and the ω_i are the weights. Because these formulas also evaluate $R(x;a)$ at just a finite number of points, x_i, quadrature-based projection techniques are essentially weighted collocation methods, but may be better because they use information at more points.

Step 5 determines a through either a minimization algorithm (in the least-squares approach) or a nonlinear equation solver applied to the system $P(a) = 0$. In what follows we illustrate some of the ways available to solve $P(a) = 0$ in a specific example.

The projection method is a general approach for the numerical solution of functional equations that arise in economic analysis. This section has presented the general framework. We next illustrate its use in specific applications in dynamic economics.

5. PROJECTION METHODS FOR TIME-HOMOGENEOUS MODELS

The various approaches to solving rational expectations models differ in three basic ways: first, in the choice of finite-dimensional approximations to functions; second, in the way the expectation in (3.4) is computed; and, third, in the method used to find an approximate solution. The work discussed here touches on two of the three critical elements – the method used to approximate equilibrium policy and pricing functions, and the method for solving the identifying conditions. In this section we focus on various combinations of approximation and solution methods that appear to be promising in the context of large multiagent rational expectations models.

5.1. Approximating Equilibrium Functions

The first key step in solving problems with heterogeneous agents is approximating the decision rules and pricing functions in an economical fashion. For example, to solve (3.4), we need to approximate two functions, $C^1(k_1, k_2, \theta)$ and $C^2(k_1, k_2, \theta)$, each of which is a function of two variables. One obvious possibility is to use polynomials. For example, we could set

$$C^i(k, \theta; a) \doteq \sum_{j_1=0}^{J_1} \sum_{j_2=0}^{J_2} \sum_{m=0}^{M} a^i_{j_1 j_2 m} k_1^{j_1} k_2^{j_2} \theta^m, \quad i = 1, 2. \tag{5.1}$$

However, ordinary polynomials are not advisable because conditioning problems (similar to multicollinearity problems in regression) make it difficult to identify the a coefficients. Judd (1992) and Judd and Gaspar (1997) instead advocated the use of orthogonal polynomials, resulting in approximations of the form

$$C^i(k, \theta; a) \doteq \sum_{j_1=0}^{J_1} \sum_{j_2=0}^{J_2} \sum_{m=0}^{M} a^i_{j_1 j_2 m} \varphi_{j_1}(k_1) \varphi_{i_2}(k_2) \psi_m(\theta), \quad i = 1, 2,$$

where $\varphi_i(.)$ ($\psi_m(.)$) is a degree $i - 1$ ($m - 1$) polynomial from some appropriate orthogonal family. For example, Chebyshev polynomials are natural to use in the k dimensions because k is expected to stay in some compact domain, but Hermitan polynomials should be used for the θ dimension because θ is a normal random variable. One could also use splines to approximate equilibrium policy functions when the number of types is small. This is discussed more extensively in the discussion of incomplete asset markets.

For problems with several kinds of agents, forms such as (5.1) suffer from a curse of dimensionality. To counter that, Judd and Gaspar (1997) advocate the use of complete polynomials. The key fact about complete polynomials is that one eliminates from (5.1) terms of high total power. In particular, a degree d approximation would use approximations of the form

$$C^i(k,\theta;a) \doteq \sum_{\substack{0 \le j_1 + j_2 + m \le d \\ 0 \le j_1, j_2, m \le d}} a^i_{j_1 j_2 m} k_1^{j_1} k_2^{j_2} \theta^m, \quad i = 1, 2.$$

In general, if there are n agents and we wanted to use a multivariate orthogonal polynomial approximation, we would use

$$C^i(k,\theta;a) \doteq \sum_{\substack{0 \le j_1 + \cdots + j_n + \ell \le d \\ 0 \le j_i, \ell \le d}} a^i_{j_1 \cdots j_n \ell} \varphi_{j_1}(k_1) \cdots \varphi_{j_n}(k_n) \psi_\ell(\theta).$$

Further simplification is possible if symmetry properties are present. For example, suppose there are three types of agents with identical preferences but different wealth. Then, type 1 agents do not care if type 2 agents are poor and type 3 agents are rich or if the reverse is true, but type 1 agents do care about the distribution. This symmetry condition imposes further conditions on the a coefficients, further reducing the number of unknowns. Specifically, the consumption function for type 1 agents has the form

$$C^1(k,\theta;a) \doteq \sum_{\substack{0 \le i + j + \ell \le d \\ 0 \le i, j, \ell \le d}} a_{j_1 \cdots j_n \ell} \varphi_i(k_1) \varphi_j(k_2, \ldots, k_n) \psi_\ell(\theta), \qquad (5.2)$$

where each $\varphi_j(k_2, \ldots, k_n)$ is a symmetric polynomial in $(k_2, \ldots, k_n)$ of total degree j. The symmetric polynomials have a particular structure, built up from a few basic symmetric polynomials. For example, the degree 1 symmetric polynomial in $(x, y, \ldots, z)$ is $x + y + \cdots + z$, the degree 2 symmetric polynomials in $(x, y, \ldots, z)$ are linear combinations of $x^2 + y^2 + \cdots + z^2$ and $(x + y + \cdots + z)^2$, and degree 3 symmetric polynomials in $(x, y, \ldots, z)$ are linear combinations of $x^3 + y^3 + \cdots + z^3$, $x^2 y + x^2 z + \cdots + y^2 z + \cdots$, and $(x + y + \cdots + z)^3$. Consumption of type m agents is defined by using the same coefficients used in (5.2) but reversing the roles of k_1 and k_m, resulting in the consumption function

$$C^m(k,\theta;a) \doteq \sum_{\substack{0 \le i + j + \ell \le d \\ 0 \le i, j, \ell \le d}} a_{j_1 \cdots j_n \ell} \varphi_i(k_m) \varphi_j(k_1, \ldots, k_{m-1}, k_{m+1}, \ldots, k_n) \psi_\ell(\theta).$$

Krusell and Smith (1997) propose a method that takes this process one step further. They examine a model with a continuum of agents. At first, it would seem impossible to use the approach in (5.2), but Krusell and Smith focus on moments. They assume that the consumption rule for any agent depends on his or her wealth (using spline approximations in the own wealth dimension) and

the moments of the distribution of wealth. This dependence on moments is a further extension of the idea of using symmetry to reduce the complexity of the approximation used for the consumption function. This is clearly seen from the definition of moments. For example, the mean capital stock is $\sum_i k_i$, which is the degree 1 symmetric polynomial in the k_i. The variance is a linear combination of $\sum_i k_i^2$ and $(\sum_i k_i)^2$, which are the degree 2 symmetric polynomials. The theory of complete polynomials tells us that the complete degree 2 approximation would consist of a linear combination of the mean, the variance, and the square of the mean. It also says that a third-order complete representation would involve the mean cubed, the third moment, and the skewness. Using their moment approach, Krusell and Smith are able to analyze how the distribution of wealth interacts with idiosyncratic and systematic risks in a real business cycle model. Surprisingly, they argue that a scheme using a few moments produces an acceptable approximation of aggregate fluctuations.

den Haan (1997) examined a similar problem where equilibrium depends on the distribution of wealth, but he takes a different approach. He approximates the distribution function with some functional family with coefficients b and then assumes that an individual's consumption depends on his or her wealth and the coefficients b that describe the distribution of wealth. This is more general than the method of Krusell and Smith, because one way to parameterize a distribution is through its moments. Even if one just focuses on the moments, there is always some implicit mapping between the moments and the distribution being used. den Haan's approach makes that mapping explicit and is more flexible than the Krusell–Smith method. For example, the den Haan method could, through appropriate parameterization, model distributions with mass points, but methods that focus on moments have greater difficulty handling mass points.

A key element of any algorithm is the manner in which equilibrium decision rules and pricing functions are approximated. We want to use a method that has few unknown parameters but is flexible, capable of approximating equilibrium with small errors. Recent papers have shown that it is important to exploit known features of equilibrium, such as symmetry, because they can drastically reduce the number of free parameters without creating an unreasonable approximation error.

5.2. Solving for the Unknown Coefficients

We next create identifying conditions for the coefficients a and solve for a. We first create some projection conditions $P(a)$, which is a finite number of conditions on the coefficients a. For example, in the case of (3.4), define the residual function to be the Euler equation errors, as in

$$R(k, \theta; a) = u_i'(C^i(k, \theta; a)) - \beta E\{u'(C^i(k^+, \tilde{\theta}; a))R(k^+, \tilde{\theta}) \mid \theta\}, \quad i = 1, 2. \tag{5.3}$$

The expectation operator in (5.3) has to be approximated, producing the approximate residual function for agent i,

$$\widehat{R}^i(k,\theta;a) = u_i'(C^i(k,\theta;a)) - \beta\widehat{E}\{u_i'(C^i(k^+,\tilde{\theta};a))F_k(k^+,\tilde{\theta}) \mid \theta\}, \quad i = 1,2,$$

where $\widehat{E}$ represents some numerical approximation of the enclosed integral. The expectation can be approximated by Monte Carlo integration, Newton–Cotes integration, or a Gaussian integration formula. Judd and Gaspar (1997) uses Gauss–Hermite, but note that a variety of integration methods, such as monomial rules and good lattice points, may be better in the multidimensional context.

The residual function is used to construct the identifying conditions. Define the projections

$$P_{ij}(a) \equiv \int_{\theta_m}^{\theta_M}\int_{k_m}^{k_M}\int_{k_m}^{k_M} \widehat{R}^i(k,\theta;a)\psi_j(k,\theta)w(k,\theta)dk_1dk_2d\theta, \quad i = 1,2,$$

where $\psi_j(k,\theta)$ are distinct functions. The projection conditions themselves are integrals that can be computed by using Monte Carlo methods, or they can be conditions motivated by orthogonal polynomial theory and Gaussian quadrature.

Once we have specified the projection, $P(a)$, we need to solve $P(a) = 0$. There are several methods available. Newton's method[4] treats the conditions $P(a) = 0$ as a system of nonlinear equations and solves for a by repeated linear approximations. Newton's method is locally quadratically convergent, but each step uses $O(n^3)$ time because it computes a Jacobian. Some refinements economize on this by approximating the Jacobian, but the computational cost per step is still a problem.

Two other procedures are motivated by economic intuition. Time iteration executes the iteration

$$\begin{aligned}\hat{C}^{i,j+1}(k,\theta) = (u')^{-1}(\beta\hat{E}\{u_i'(\hat{C}^{i,j}(Y(k,\theta) - \hat{C}^{i,j+1}(k,\theta),\tilde{\theta}))\\ \times F_k(Y(k,\theta) - \hat{C}^{i,j}(k,\theta),\tilde{\theta}) \mid \theta\}).\end{aligned} \tag{5.4}$$

For a fixed (k,θ) vector, (5.4) is a nonlinear equation in $\hat{C}^{i,j+1}(k,\theta)$. Solving (5.4) for several choices of (k,θ) generates values for $\hat{C}^{i,j+1}(k,\theta)$, information that is then used to compute the coefficients for $\hat{C}^{i,j+1}(k,\theta)$. Time iteration corresponds to solving the corresponding dynamic program problem backward in time.

Successive approximation methods proceed more directly, using less computation per step. Specifically, successive approximation takes the policy functions

[4] It is well known that one should not apply the original Newton method. It is more advisable to use an implementation of Powell's hybrid method, such as that contained in the MINPACK collection. One could also use the more advanced TEN-SOLVE package.

computed in iteration j, $\hat{C}^{i,j}$, and applies the computation

$$\hat{C}^{i,j+1}(k,\theta) = (u')^{-1}(\beta \hat{E}\{u'(\hat{C}^{i,j}(Y(k,\theta) - \hat{C}^{j}(k,\theta),\tilde{\theta})) \times F_k(Y(k,\theta) - \hat{C}^{i,j}(k,\theta),\tilde{\theta})|\theta\}) \tag{5.5}$$

at a finite number of points (k,θ) to produce $\hat{C}^{i,j+1}(k,\theta)$ data sufficient to fix the unknown coefficients of $\hat{C}^{i,j+1}$.

Both successive approximation and time iteration are only linearly convergent. Because $\hat{C}^{i,j+1}(k,\theta)$ is expressed directly in terms of the right-hand side of (5.5), the computation cost is smaller for successive approximation. Successive approximation was used in the rational expectations by Miranda and Helmburger (1988), who observed that it was an efficient method for computation. It can also be motivated by learning arguments in Marcet and Sargent (1989). Successive approximation is often quite stable, converging to the equilibrium, and the computational cost of each iteration is only $O(n^2)$. For the case of a simple growth problem, Judd (1998, pp. 557–558) shows that successive approximation is locally convergent except for some extreme choices of tastes and technology. Time iteration is more reliable but generally slower than successive approximation when the latter converges. Time iteration was used by Gustafson, and in the Wright and Williams work, and theory indicates that it will be much slower than Newton's method for small problems and slower than successive approximation. However, the performance of successive approximation in more general, multiagent and multigood contexts is unclear. The critical fact is that successive approximation methods, like all Gauss–Seidel style methods, need stability in all dimensions (see the discussion of Gauss–Seidel methods for linear equations in Judd, 1998), and as one adds dimensions, it is more likely that the additional dimensions include at least one that causes instability. In contrast, the strong connection between value function iteration and time iteration of Euler equations indicates that time iteration is more robust.

The time iteration method is also used by Rios-Rull (1999), who solves for individual value functions as well as policy functions in recursive equilibria. Sometimes equilibrium is best expressed in terms of individual value functions. See Rios-Rull (1999) for a detailed presentation of that approach. Of course, Newton's method or similar nonlinear equation methods could also be used to solve problems formulated in terms of value functions, because the equilibrium is approximated by a nonlinear set of conditions on the coefficients of the parameterization of the value and policy functions.

6. INCOMPLETE ASSET MARKETS WITH INFINITELY LIVED AGENTS

The incomplete asset market model with infinitely lived agents is one that has been analyzed using computational methods in several recent papers. We discuss the model in some detail in order to show how the ideas from Section 5 can be applied to a specific economic environment. We give a simple example

to highlight the critical numerical issues that arise in asset market problems. We also review some alternative approaches to the approximation of equilibria in this model.

6.1. A Model of Incomplete Asset Markets

Consider a Lucas (1978) economy with heterogeneous agents, a single commodity, and incomplete asset markets. There are H infinitely lived investors. At each period $t = 0, 1, \ldots$, investor h receives a stochastic labor income e_t^h. In addition there is a Lucas tree (which we refer to as the stock) with stochastic dividends d_t. At $t = 0$ each agent h owns a fraction of the tree $s_{-1}^h \geq 0$. The tree is in unit net supply, $\sum_h s_{-1}^h = 1$, so that aggregate endowments (output) at each time t equal $\sum_{h=1}^H e_t^h + d_t$. All uncertainty can be described by a time-homogeneous finite-state Markov process. Let $Y = \{1, 2, \ldots, S\}$ denote the exogenous states, and y_t be the time t value. Individual labor endowments $e^h : Y \to \mathbb{R}_{++}$ and the dividends $d : Y \to \mathbb{R}_+$ depend on the current state y alone.

Each agent h maximizes the expected utility function

$$U_h(c) = E\left\{\sum_{t=0}^{\infty} \beta^t u_h(c_t, y)\right\},$$

over possible infinite consumption streams c. We assume that the utility functions $u_h(., y) : \mathbb{R}_{++} \to \mathbb{R}$ are strictly monotone, C^2, strictly concave, and satisfy $\lim_{x\to 0} u_c(x, y) = \infty$. We also assume that the discount factor $\beta \in (0, 1)$ is the same for all agents, and that agents have common beliefs about the transition matrix for the exogenous states.

Agents trade two securities in order to smooth their consumption across time and states. They can trade shares of equity denoting ownership in the Lucas tree and a one-period bond in every time period. One bond at time t delivers one unit of the consumption good at time $t + 1$ for any y_{t+1}. Bonds are in zero net supply. Markets are incomplete if $S > 2$ and perfect risk sharing will generally be impossible. Let b^h denote an agent's bond holding, s^h his or her stock holding, q^s the price of the stock, and q^b the price of the bond. At each time t agent h faces the budget constraint

$$c_t^h = e^h(y_t) + b_{t-1}^h + s_{t-1}^h\left(q_t^s + d(y_t)\right) - b_t^h q_t^b - s_t^h q_t^s.$$

In addition, we assume the short-sale constraints

$$b_t^h \geq K_b^h \text{ and } s_t^h \geq K_s^h, \quad \forall h = 1, \ldots, H,$$

where $K_b^h < 0$ and $K_s^h \leq 0$. As we will see, these last constraints are needed to obtain the existence of an equilibrium, but they turn out to present special challenges for any computational strategy.

6.2. Recursive Equilibria

It is well known that the model always has a competitive equilibrium, that is, there exist prices and allocations such that all markets clear and all agents maximize utility subject to their budget restrictions (see, e.g., Magill and Quinzii, 1996). However, to compute an equilibrium for an infinite-horizon model, one must focus on recursive equilibria. Recursive equilibria are dynamically simple, expressing prices, trades, and consumption as a time-invariant function of a finite number of state variables. In this problem, the state variables include the exogenous states y and the agents' portfolios. This problem is much more difficult than the original Lucas (1978) model, where all equilibrium values depend solely on the exogenous state. Because of agent heterogeneity, the state space includes the portfolios because the distribution of wealth will influence equilibrium prices. For the incomplete asset model, it is standard to assume that the exogenous income and dividend state $y \in Y$ together with the agents' portfolio holdings $\Theta := (b^h, s^h)_{h=1}^H$ constitute a sufficient state space. Although there are examples of economies where recursive equilibria do not exist (see Kubler and Schmedders, 2002), we proceed under this assumption.

We denote the endogenous state space of all possible portfolio holdings of all agents by Z. Because of the short-sale constraints, the set Z is compact.[5] Furthermore, we assume that the recursive equilibrium can be described by a family of continuous policy functions $f^h = (f^{hb}, f^{hs}) : Y \times Z \to Z$, $h = 1, \ldots, H$, which determine agents' optimal portfolio choices given portfolio holdings and the income state of the current period, and by a continuous price function $g = (g^b, g^s) : Y \times Z \to \mathbb{R}^2_{++}$, which maps the current state into the asset prices q^b and q^s.

The equilibrium functions f and g are defined by the following requirements.

1. (RE1): Market clearing:
$$\sum_{h=1}^{H} f^{hb}(y, \Theta) = 0, \ \sum_{h=1}^{H} f^{hs}(y, \Theta) = 1, \quad \forall y \in Y, \Theta \in Z.$$
2. (RE2): Consumption choices are consistent with wealth and asset trades for all $y \in Y$ and all $\Theta \in Z$:
$$\begin{aligned} c^h &= c^h(y, \Theta) \\ &= e^h(y) + b^h + s^h(g(y, \Theta) + d(y)) - f^h(y, \Theta) g(y, \Theta). \end{aligned}$$
3. (RE3): Choices are optimal; hence, for any two subsequent exogenous

[5] In fact, short-sale constraints imply that
$$Z = \prod_{h=1}^{H} \left(\left[K_b^h, -\sum_{i \neq h} K_b^i \right] \times \left[K_s^h, 1 - \sum_{i \neq h} K_s^i \right] \right).$$

states y and y_+ and all $\Theta \in Z$, consumption satisfies

$$
\begin{aligned}
&c_+^h := c^h(y_+, f^h(y, \Theta)),\\
&q_+^s := g^s(y_+, f^h(y, \Theta)),\\
&\lambda^{hb} := u_h'(c)g^b(y, \Theta) - \beta E(u_h'(c_+^h)) \geq 0,\\
&\lambda^{hs} := u_h'(c)g^s(y, \Theta) - \beta E(u_h'(c_+^h)(q_+^s + d(y_+))) \geq 0,\\
&\lambda^{hs}\left(f^b(y, \Theta) - K_b^h\right) = 0,\\
&\lambda^{hb}\left(f^s(y, \Theta) - K_s^h\right) = 0.
\end{aligned}
$$

The macroeconomic literature often assumes stationary growth for endowments. If all agents have identical constant relative risk-aversion utility, (RE1)–(RE3) can be rewritten in terms of consumption–wealth ratios, transforming the nonstationary growth problem into a problem confined to a compact set of ratios.

6.3. Kuhn–Tucker Conditions as a System of Equations

As a result of the short-sale constraints, the agents face utility maximization problems with inequality constraints, resulting in first-order conditions of optimality that include shadow prices and inequalities; see the equations in (RE3). By means of a simple trick (see Garcia and Zangwill, 1981), Judd, Kubler, and Schmedders (1999b) transform the collection of equations and inequalities into a nonlinear system of equations, which can be solved using a nonlinear-equation routine.

Let l be a natural number and $\alpha^{ha} \in \mathbb{R}$ for $h = 1, 2$ and $a \in \{b, s\}$. Note the following relations:

$$
(\max\{0, \alpha^{ha}\})^l = \begin{cases} (\alpha^{ha})^l & \text{if } \alpha^{ha} > 0,\\ 0 & \text{if } \alpha^{ha} \leq 0 \end{cases}
$$

$$
(\max\{0, -\alpha^{ha}\})^l = \begin{cases} 0 & \text{if } \alpha^{ha} > 0.\\ |\alpha^{ha}|^l & \text{if } \alpha^{ha} \leq 0 \end{cases}
$$

Moreover,

$$
(\max\{0, \alpha^{ha}\})^l \geq 0,\ (\max\{0, -\alpha^{ha}\})^l \geq 0,\ (\max\{0, \alpha^{ha}\})^l(\max\{0, \alpha^{ha}\})^l = 0.
$$

We define

$$
\lambda^{ha} = (\max\{0, \alpha^{ha}\})^l, \quad f^a(y, \Theta) - K_a^h = (\max\{0, -\alpha^{ha}\})^l,
$$

which allows us to state the first-order conditions of optimality as a system of equations that is equivalent to those of (RE3):

$$
\begin{aligned}
&-g^b(y, \Theta)u_h'(c) + \beta_h E_t(u_h'(c_+^h))) + (\max\{0, \alpha^{hb}\})^l = 0, \qquad \text{(RE3}'\text{)},\\
&-f^b(y, \Theta) + K_b^h + (\max\{0, -\alpha^{hb}\})^l = 0,\\
&-g^s(y, \Theta)u_h'(c) + \beta_h E_t\{(q_+^s + d(y_+))u_h'(c_+)\} + (\max\{0, \alpha^{hs}\})^l = 0,\\
&-f^s(y, \Theta) + K_s^h + (\max\{0, -\alpha^{hs}\})^l = 0.
\end{aligned}
$$

We have thus transformed the optimality conditions into a system of equations that can be solved with standard numerical techniques.

6.4. Bounded Portfolio Space

The computational challenge is to approximate the equilibrium functions f and g. At this point, it is important to stress that the short-sale constraints are essential for the computations. From an economic modeling perspective, short-sale constraints are an undesirable part of the model. The bounds on short sales have to be chosen exogenously, and because in reality explicit short-sale constraints rarely exist, this choice cannot be guided by data. Although economic agents do face trading restrictions and debt constraints, there are often no legal limits on short positions in individual securities.

From a theoretical point of view, to close the model, one must rule out Ponzi schemes, that is, the possibility of an infinite accumulation of debt. The standard approach (see, e.g., Levine and Zame, 1996) is to impose an implicit debt constraint. For a model with only a single bond, implicit debt constraint can often be reformulated as a short-sale constraint. Zhang (1997a, 1997b) develops algorithms to compute equilibria in models with a single asset and implicit debt constraints. Unfortunately, his approach does not generalize to models with more than one asset. In particular, competitive equilibria do not always exist because agents can inflate their portfolios without any bounds without violating the implicit debt constraint.

For all practical purposes, it is therefore crucial to ex ante fix a bounded set of admissible portfolio holdings for all agents. The easiest way to obtain a bounded set of portfolios is to impose a short-sale constraint. Short sales can be constrained through a priori specified fixed exogenous lower bounds on the portfolio variables. In equilibrium, when all financial markets are required to clear, all agents' portfolios are also bounded above, resulting in a compact set of admissible portfolios for the entire economy. Note that we define the bounds on short sales as agent dependent because it is certainly realistic to assume that an agent's income influences how much he or she can borrow.

Judd et al. (1999a) find that, in many cases, when there are two assets, short-sale constraints will frequently be binding.

6.5. Computational Errors

When approximating policy and price functions, one has to deal with various kinds of computational errors, and it is impossible to determine the equilibrium functions exactly. This fact of life leads to the central question of how large an error is acceptable and when to stop an approximation algorithm. The typical procedure used to solve our model is of an iterative nature and terminates when a stopping rule is satisfied. Such stopping rules do not specify when the approximate equilibrium prices are close to the true equilibrium prices, but instead when the difference between consecutive approximations is small.

The method is then thought to have stabilized around an approximate solution. At this point, however, it remains unclear how close the computed prices and portfolios are to the true equilibrium prices and portfolios and if we actually have an approximate description of a recursive equilibrium. There is an obvious need for a close evaluation of the computed solutions.

Ideally, one wants to derive error bounds or accuracy estimates of the computed solutions. Although bounds like this exist for finite-dimensional problems (see, e.g., Blum et al., 1998) and for numerical dynamic programming (see Santos and Vigo-Aguiar, 1998), there exist no comparable theories for the general equilibrium model under consideration.

A popular approach for verifying the quality of a solution is to compute the maximum relative errors in the agents' first-order conditions. For the case in which short-sale constraints are not binding and the related shadow prices are zero, conditions (RE3) imply that the maximum relative errors are given by

$$\max_{\theta} \left\| \frac{\beta E(u_h'(c_+^h)) - u_h'(c) g^b(y, \Theta)}{u_h'(c) g^b(y, \Theta)} \right\|$$

and

$$\max_{\theta} \left\| \frac{\beta E((q_+^s + d(y)) u_h'(c_+^h)) - u_h'(c) g^s(y, \Theta)}{u_h'(c) g^s(y, \Theta)} \right\|.$$

Unfortunately, low errors in agents' Euler equations do not give any indication of how close we are to an equilibrium. This well-known fact has various interpretations in the literature. Judd (1992) argues that it is not sensible to expect infinite precision from agents and that therefore the computed prices and allocations are likely to be a good description of the actual economic outcome. For this line of reasoning it is important to show how small the errors actually are. Without knowing the actual solution, this is not unambiguously possible. Judd (1992) suggests evaluating the Euler equations at the computed prices and allocation and computing the wealth equivalent of the Euler equation residual when projected in directions not used to compute the approximation. A small error here would be consistent with the interpretation of an approximate equilibrium in the sense that agents are close to rational.

Heaton and Lucas (1996) and Telmer (1993) use a slightly different error criterion. For each agent, they compute the asset prices that support the agents' computed decision. When there are two agents, they therefore get prices q_1^s, q_1^b for the first agent and q_2^s, q_2^b for the second agent. They then report $(q_1^s - q_2^s)/q_1^s$ and $(q_1^b - q_2^b)/q_1^b$.

When the short-sale constraint is not binding, these two formulations are very similar but not identical. In general, q_2 should provide a good approximation for the equilibrium price. Under q_2, relative errors in the Euler equation for the bond (the analysis for the stock is analogous) are $\{q_2^b u_1'(c) - \beta E[u_1'(c_+)]\}/[q_2^b u_1'(c)]$.

By definition of q_1^b, we have

$$\frac{q_2^b u_1'(c) - \beta E(u_1'(c_+))}{q_2^b u_1'(c)} = \frac{(q_2^b - q_1^b)u_1'(c)}{q_2^b u_1'(c)} = \frac{q_2^b - q_1^b}{q_2^b}.$$

The errors in both agents' Euler equations can presumably be decreased by setting $q^* = (q_1 + q_2)/2$; the errors reported in Heaton and Lucas and Telmer are therefore likely to slightly overstate the errors in the Euler equation. In most cases, however, the difference will be negligible.

7. A SPLINE COLLOCATION ALGORITHM

Here we want to argue that it is sensible to approximate equilibria with two assets over a continuous endogenous state space. For such a continuous approximation, a family of polynomials is used to approximate these functions. Here we review the spline collocation method used in Judd et al. (2000). In Section 9.2 we examine the method used in Marcet and Singleton (1999) and discuss methods that discretize the endogenous state space.

There are two issues one has to face when developing an algorithm to approximate recursive equilibria. First, one has to find a scheme to approximate the true equilibrium functions f and g. Generally, the approximating functions will be determined by a finite number of parameters. The second step must then be to solve for these unknown parameters.

Judd et al. (2000) use cubic splines to approximate the equilibrium functions and compute the spline coefficients using collocation methods. They solve the collocation equations with an iterative approach. We briefly describe their algorithm and use a simple example to show where the difficulties lie in approximating and in solving for the equilibrium functions.

The main steps of the algorithms are as follows:

1. (C1): The equilibrium functions f and g are approximated by piecewise polynomial functions. They can therefore be parameterized by a finite number of coefficients.
2. (C2): In each set of the endogenous state space $[K_b^1, -K_b^2] \times [K_s^1, -K_s^2]$, choose a finite number of $N \times N$ points, called *collocation points*.
3. (C3): The algorithm searches for coefficients of the approximating functions that ensure that, at the collocation points, the Euler equations and market-clearing conditions hold.

7.1. Representing the Equilibrium Functions

The equilibrium functions are defined on an uncountable set, and so we cannot specify all their values exactly. Instead, it is feasible for us only to approximate the functions f and g using a fairly small number of parameters. As we show in what follows, these functions are likely to have extremely high curvature in certain regions. In fact, if the short-sale constraints are modeled explicitly,

these equilibrium functions fail to be C^1. A global polynomial approximation (using, e.g., orthogonal polynomials) therefore does not work unless we are willing to allow for large computational errors. The most sensible approach turns out to be to approximate the equilibrium function in a piecewise fashion by finitely parameterized functions $\hat{f}$, $\hat{g}$, using relatively few parameters. We use piecewise cubic polynomials (cubic splines) to approximate them.

One-dimensional cubic splines are easily defined. Given m points of a real-valued function $(x_i, y_i)_{i=1}^m$, a cubic spline $s(x)$ is defined by the requirement that $s(x_i) = y_i$ for all $i = 1, \ldots, m$, that in each interval $[x_i, x_{i+1}]$ the function s is a cubic polynomial, and that s is C^2 on $[x_1, x_m]$. By representing them as a linear combination of a collection of so-called B splines, we can use splines for approximating higher-dimensional functions. See de Boor (1978) for a thorough description of splines and their approximating properties.

Note that, by approximating the equilibrium functions by means of piecewise linear combinations of polynomials, we have transformed the equilibrium computation from finding function values on an uncountable set to finding finitely many weights for the appropriate linear combinations. Put differently, the problem has been reduced from computing infinitely many values to finding a reasonable finite number of parameters.

7.2. A Time-Iteration Algorithm

With the use of the collocation method to find approximating functions $\hat{f}$ and $\hat{g}$, the crucial problem has now become to solve for the spline coefficients. To obtain sufficient accuracy, we find that the number of unknown coefficients turns out to be rather large (several thousand). This large number of parameters makes it difficult for us to solve for them directly; we apply an iterative approach instead.

The basic intuition is that, at each iteration i, we take next period's policy functions as given and compute at every collocation point this period's portfolio holdings and prices that satisfy the Euler equation. Given functions $\hat{f}_i$ and $\hat{g}_i$, we obtain $\hat{f}_{i+1}$ and $\hat{g}_{i+1}$ by interpolating the computed portfolio holdings and prices at the collocation points. Recursive infinite-horizon equilibria are approximated by finite-horizon equilibria as the number of periods becomes very large. For discount factors β close to one, the number of iterations needed to obtain a satisfactory approximation will be very large (for $\beta = 0.99$ it can lie around 200). It is therefore important that, in each iteration, an efficient way is found to solve the Euler equations. We come back to this problem in the context of our example.

The algorithm can be summarized as follows, as a time-iteration spline collocation algorithm.

Step 0: Select a set G of collocation points and a starting point $\hat{f}_0$, $\hat{g}_0$.
Step 1: Given functions $\hat{f}_i$, $\hat{g}_i$, $\forall \theta \in G$ and $\forall y \in Y$, compute portfolios and prices solving the Euler equations and market-clearing conditions.

Step 2: Compute the new approximations $\hat{f}_{i+1}, \hat{g}_{i+1}$ by means of interpolation.
Step 3: Check stopping criterion: If the errors are sufficiently small, then go to Step 4. Otherwise, increase i by 1 and go to Step 1.
Step 4: The algorithm terminates. Set $\hat{f} = \hat{f}_{i+1}$ and $\hat{g} = \hat{g}_{i+1}$; these are the approximate equilibrium functions.

8. THE MAIN COMPUTATIONAL CHALLENGES AND AN EXAMPLE

As it turns out, approximating equilibria for models with more than one asset is a difficult task. Here we use a simple example to illustrate the following two problems.

1. Short-sale constraints are frequently binding, resulting in nonsmooth policy functions. Global polynomial approximation schemes cannot be used to approximate these functions.
2. Even in regions of the state space where constraints are not binding, the system of equalities describing equilibrium is extremely ill conditioned. Without a very good starting point, Newtonian methods cannot find a solution.

There are two investors with identical constant relative risk aversion (CRRA) utility, $u_h(c) = (c^{1-\gamma})/(1-\gamma)$ with a coefficient of relative risk aversion of $\gamma = 1.5$ and a discount factor $\beta = 0.96$. There are four exogenous states, idiosyncratic shocks to labor income are

$$e^1 = \begin{bmatrix} 3 \\ 3 \\ 7 \\ 7 \end{bmatrix}, \quad e^2(y) = 10 - e^1(y),$$

and the stochastic dividends – the only source of aggregate uncertainty in our simple example – are given by $d = (2.5, 2, 2.5, 2)'$.

The transition matrix Π is chosen to ensure that idiosyncratic shocks are very persistent. This seems to be a stylized fact and, as we shall see, it causes substantial computational difficulties. Here Π is given by

$$\Pi = \begin{bmatrix} 0.4 & 0.4 & 0.1 & 0.1 \\ 0.4 & 0.4 & 0.1 & 0.1 \\ 0.1 & 0.1 & 0.4 & 0.4 \\ 0.1 & 0.1 & 0.4 & 0.4 \end{bmatrix}.$$

8.1. Short-Sale Constraints

As explained in Subsection 6.4, short-sale constraints are an essential part of our model. We set the short-sale constraints to be $K_b^h = -2.5$ and $K_s^h = -0.1$

for both agents $h = 1, 2$. Figure 7.1 depicts investor 1's computed equilibrium demand for bonds, $\hat{f}_y^{1b}$, for the exogenous shock of $y = 1$ for our example. For large regions of the endogenous state space, the short-sale constraint on the bond is binding. The economic interpretation of this is straightforward: Investors cannot use the stock to insure against bad idiosyncratic shocks ex ante. Given a bad shock, the only possibility to smooth consumption is therefore to borrow. Such can be achieved by selling either the stock or the bond. However, for most levels of stock holdings, the investor prefers to sell the bond and does so up to the short-sale constraints. Although the given parameterization is a very stylized example, this phenomenon is very likely to occur in models with realistically calibrated persistent idiosyncratic shocks.

The main computational problem with short-sale constraints is the fact that the resulting policy functions f are no longer smooth. Clearly, the plotted $\hat{f}_1^{1b}$ does not approximate a smooth function. At those values of last period's portfolio holdings where the short-sale constraint becomes binding, the policy function is not differentiable. This fact shows why it is impossible to approximate these functions globally with polynomials. Even with splines, a good approximation is possible only if the number of collocation points is very high; in this example we chose 20×20 collocation points. Of course, by the definition of cubic splines, the approximating function is still C^2 while the true function is not. In fact, one can see in Figure 7.1 that the approximating function

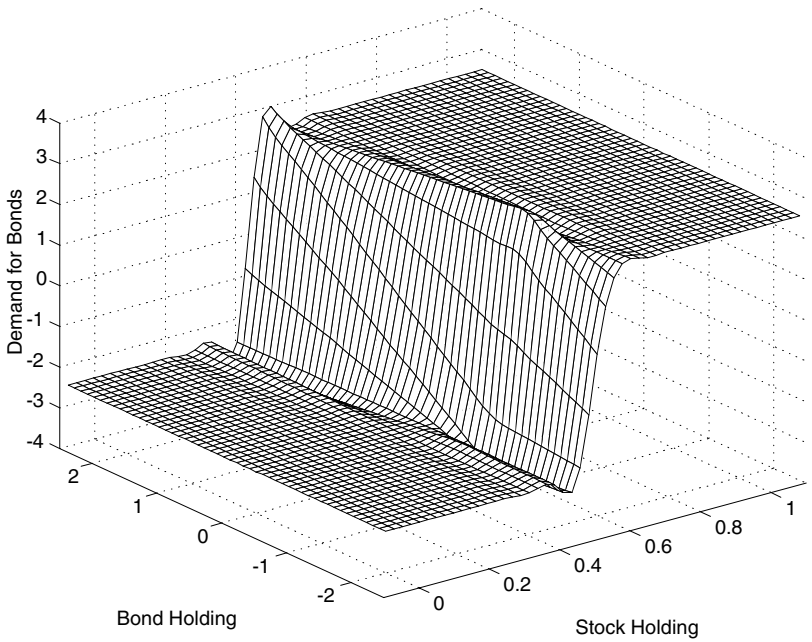


Figure 7.1. Bond demand: no transaction costs.

is slightly "hump shaped" close to the "theoretical nondifferentiability." This hump leads to substantial errors in the Euler equations. Although the resulting average errors are very low (around 0.0001 percent), the maximum error in the Euler equation lies around 0.3 percent. Furthermore, if only 225 collocation points are used, then the maximum error jumps up to 0.9 percent.

For many applications, particularly if there are more than two endogenous states, it is impractical to have twenty collocation points in each dimension. Therefore it is useful to consider variations of the model that lead to less extreme shapes of the policy functions.

8.1.1. Transaction Costs

One easy way to avoid extreme trading is to assume that trading the stock and the bond is costly. In fact, it might even be realistic to assume that there is a real cost in acquiring financial assets. The following specification of transaction costs is from Heaton and Lucas (1996, Section IV-D).

At each date t, an agent h pays transaction costs of $\omega(\theta_{t-1}^h, \theta_t^h)$. We assume that ω has the functional form

$$\omega(\theta_{t-1}, \theta_t) = \tau^b b_t^2 + \tau^s \left(q_t^s (s_t - s_{t-1})\right)^2,$$

where τ^b, τ^s are constants. The assumption of strictly convex costs is unrealistic, but it is needed to ensure that agents face a differentiable and convex programming problem. We set $\tau^s = 0.0001$ and $\tau^b = 0.0001$. Note that these costs are so small that they are not likely to have a huge effect on agents' welfare. For example, trading 2.5 units of the bond (the biggest possible amount) will cost 0.000625 percent of aggregate endowments. Costs this low are likely to have small effects on equilibrium prices. However, they substantially affect equilibrium trades. Figure 7.2 is the analog of Figure 7.1 for the case of small transaction costs. The policy function is now much better behaved. Not surprisingly, the errors are much smaller. Even with 10×10 collocation points, maximum errors lie around 0.0001 percent.

Figures 7.3 and 7.4 show $\hat{g}^b$ and $\hat{g}^s$ for the case with transaction costs. They are visually indistinguishable from the case of no transaction costs, indicating that transaction costs have negligible effects on asset prices. In fact, in simulations first and second moments of returns turn out to be within 0.01 percent of each other.

8.1.2. Penalties on Portfolios

An alternative approach that allows us to drop the assumption of short-sale constraints altogether is to assume that agents are allowed to hold portfolios of any size but get penalized for large portfolio holdings. The intuition behind such a model assumption is that there are costs associated with large short positions, and in a simplification we model them as penalties to agents' utilities; when these penalties get sufficiently large, agents avoid extreme positions. The

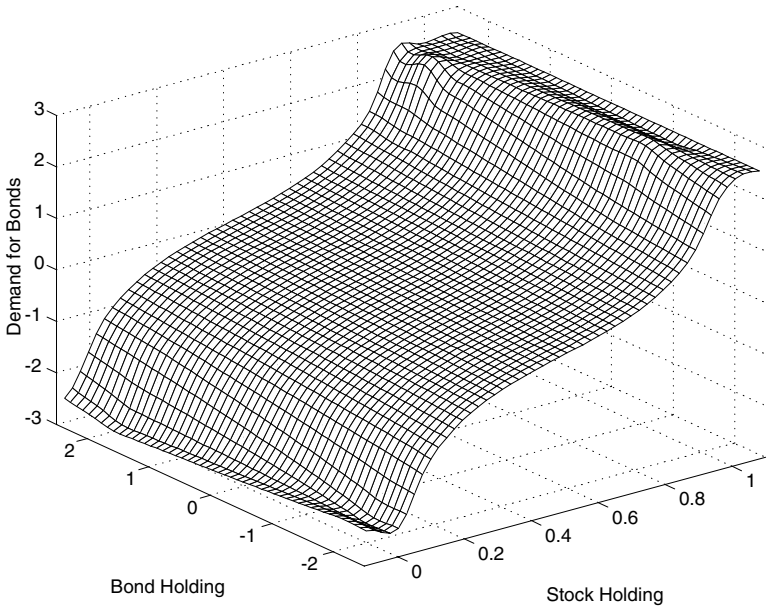


Figure 7.2. Bond demand: with transaction costs.

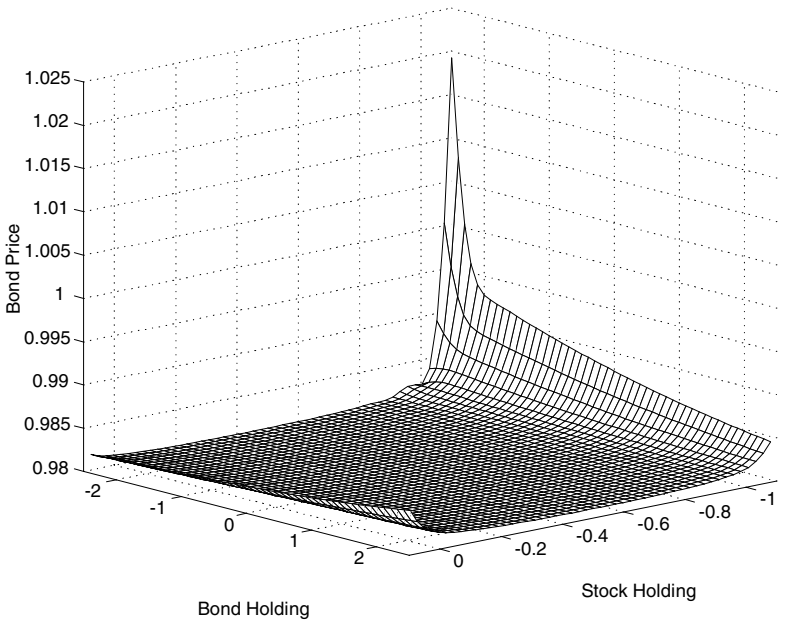


Figure 7.3. Bond price manifold.

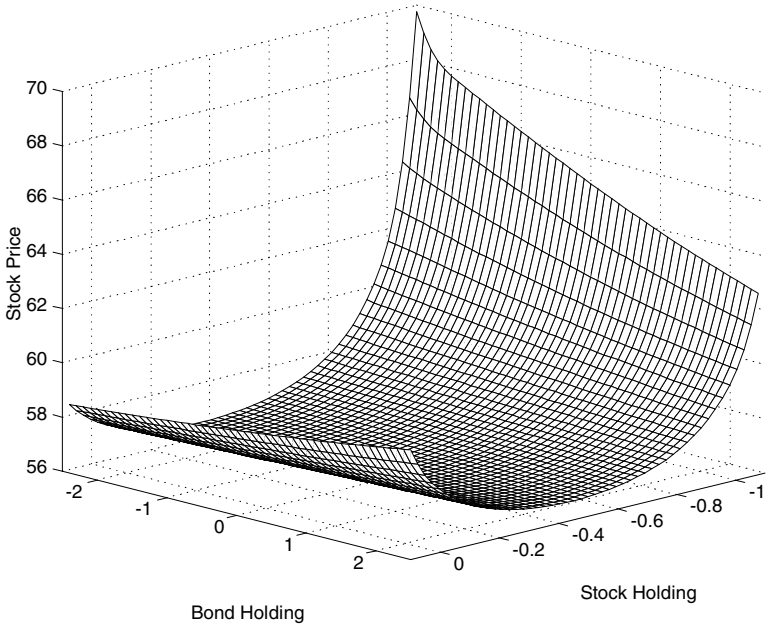


Figure 7.4. Stock price manifold.

advantage of utility penalties on large short positions is that this restriction does not constitute an a priori exogenous constraint on short sales. Instead, the penalties lead to endogenous avoidance of short sales, depending on how much agents desire large short positions.

We use a penalty function of the form

$$\rho^h(s, b) = \kappa^b \min(0, b - L^{hb})^4 + \kappa^s \min(0, s - L^{hs})^4,$$

where $\kappa^a \geq 0, a \in \{b, s\}$ and $L^{ha} \leq 0$. Note that there is no punishment for large long positions. If κ^a is sufficiently large, the penalty function almost acts like a hard short-sale constraint on the corresponding asset $a \in \{b, s\}$. For a more general description of the model, it suffices that ρ is a convex function satisfying $\rho(b, s) \to \infty$ as $b \to -\infty$ or $s \to -\infty$. The portfolio penalties lead our agents to have utility functions over consumption and portfolio holdings of the form

$$V_h(c, b, s) = U_h(c) - E\left\{\sum_{t=0}^{\infty} \beta^t \rho^h(b_t, s_t)\right\}.$$

The main problem with this approach is that the portfolio penalties ρ^h have to be chosen a priori, to guarantee that the resulting policy function maps into the endogenous state space Z. The exact choice of these penalties influences the computed equilibrium prices substantially.

8.2. Ill-Conditioned Systems

The system of Euler and market-clearing equations that define a recursive equilibrium, (RE1)–(RE3), tends to be numerically very unstable. We measure the numerical stability of a system by computing its condition number. The condition number of an invertible real $n \times n$ matrix A is defined as $\kappa(A) = \|A\|\|A^{-1}\|$, where $\|A\|$ is the operator norm,

$$\|A\| = \sup_{x \neq 0} \frac{\|Ax\|}{\|x\|}.$$

It is a useful measure of a matrix being nearly singular and is used to measure the relative error of solutions to the linear problem $Ax = b$ (see Judd, 1998, or Blum et al., 1998).

A large condition number implies that a system is sensitive to small changes and is difficult to solve. Newton's method iterates on $x_{k+1} = x_k(Df(x_k))^{-1} f(x_k)$, and a large condition number of the Jacobian $Df(x_k)$ implies that Newton's method cannot be used if one does not have a starting point very close to the solution. On normal computers with machine precision around 10^{-15}, condition numbers above 10^{10} are considered unacceptable because it indicates that an input or round-off error of ϵ leads to an output error of $10^{10}\epsilon$.

In models in which the assets have strongly correlated returns, the choices are nearly indeterminate and the condition number of the Euler equations will be very large. In dynamic models with infinitely lived agents and transitory shocks, the choice between stock and bonds is usually uniquely determined but of second-order importance to the investor. Because shocks are transitory, a bad idiosyncratic shock can generally be smoothed out by borrowing. Agents are therefore primarily interested in their total debt level, $s_t^h q_t^s + b_t^h q_t^b$. In regions where the short-sale constraint is not binding, they are nearly indifferent between stocks and bonds. In our example, the condition number of the equilibrium system at the solution lies around 10^9; in some regions of the state space it even reaches 10^{10}. This causes substantial numerical difficulties.

8.2.1. Transaction Costs

If, in our example, we impose a substantial trading cost for the stock, the conditioning of the system improves drastically. For example, if we set $\tau^s = 0.05$, the condition number decreases to around 10^6. However, a large transaction cost clearly also has a substantial impact on trading and prices. An alternative is to impose transaction costs on both bonds and stocks. With the just-given specification of $\tau^s = 0.0001$ and $\tau^b = 0.0001$, the condition number of the system decreases to 10^7.

8.2.2. Solving the Euler Equations With Homotopy Methods

In many applications, we are interested in equilibria that result without any restrictions on transactions. In this case, special care is needed to solve the

Euler equations. If we do not have a good starting point, algorithms based on Newtonian methods for solving (RE3′) are not likely to perform well because they are not globally convergent and because the system of equations is not well conditioned for many values of the endogenous variables. In this case we have to use homotopy methods to solve (RE3′). The key insight for solving system (RE3′) is that it is similar to the equilibrium conditions of the well-known General Equilibrium Model with Incomplete Markets (GEI Model). Therefore, in order to solve system (RE3′) we can apply – with some modifications – algorithms that have been developed for the GEI Model (in particular, Schmedders, 1998).

The main idea of homotopy methods is to deform a system of equations into a simple system that can be easily solved. Then this easy system is continuously transformed back into the original system. Beginning with the known solution of the easy system, a path of solutions to the encountered intermediate systems is followed, leading eventually to a solution of the given system of equations. Eaves and Schmedders (1999) give an intuitive description of the homotopy principle, addressing many issues in the framework of simple economic examples.

In the context of problems involving endogenous portfolio choices, it is often useful to set up the easy system such that the agents are forced to hold a prespecified portfolio. Such a system can easily be achieved by adding a portfolio penalty term to the agents' utility functions. In the easy system, this penalty has maximal force inducing the agents to hold the specified portfolio. As the penalty is relaxed and driven to zero, a path of portfolio choices leads to the agents' equilibrium portfolios.

9. ALTERNATIVE APPROACHES

A crucial feature of every approximation method is the way it deals with the endogenous state space Z. Here we first discuss algorithms that discretize the set Z and allow only finitely many values of the endogenous state variable. Subsequently we present the parameterized expectations algorithm of Marcet and Singleton (1999), which uses a continuous state space as the spline collocation algorithm of Section 8.

9.1. Discrete State Space

The papers by Telmer (1993), Lucas (1994), and Heaton and Lucas (1996) are examples of papers that approximate recursive equilibria by discretizing the endogenous state space Z; that is, agents' portfolio holdings can take values only in a prespecified finite set. We describe the basic ideas of these methods in the context of Lucas' (1994) two-investor two-asset model.

1. (D1): Under the assumption that there are only two investors, market clearing implies that the endogenous state space reduces to

$[K_b^1, -K_b^2] \times [K_s^1, -K_s^2]$. In each set of the product, choose a finite number of N points. Thus, the continuous two-dimensional state space has been collapsed to N^2 points.

2. (D2): The agents must choose their portfolios to always be exactly one of these N^2 points. Put differently, given an "old" portfolio in the discrete state space and a new exogenous shock $y \in Y$, the agents must choose a "new" portfolio that also lies in that discrete set. Therefore, the equilibrium policy function f^h, where $h = 1, 2$, can be represented as S collections of N^2 pairs of portfolio points in the discrete set, one collection for every exogenous state $y \in Y$. Similarly, the price functions g are also only a collection of SN^2 points.
3. (D3): The algorithm searches for asset prices and an allocation of assets that comes as close as possible to agents' optimality. The algorithm considers only those portfolio combinations of the agents that satisfy the market-clearing equation. As a result of the discretization of the state space, it is impossible that all agents' decisions are optimal. The goal of the algorithms must be to find functions g and f that minimize the errors in the agents' Euler equations or equivalently minimize the relative difference between supporting prices.

The discrete methods have the advantage that, because of their simplification of the equilibrium solution problem, they are numerically stable and easy to implement. For applications with only a single state variable (e.g., versions of the model where either the bond market or the stock market is shut down), they perform very well. However, we explain in what follows that, for models with more than one endogenous state variable, these methods are generally too slow or of far too low a precision to be effective tools.

The distinguishing feature of discrete methods is how they solve the agents' Euler equations, that is, how they perform Step (D3) of the basic approach setup. We discuss this point now in more detail.

9.1.1. A Single Security: No Equity

Telmer (1993) considers a greatly simplified version of our model and assumes that equity is not tradable; instead, agents are active only in the bond market. The state space then simplifies to the one-dimensional interval $[K_b^1, -K_b^2]$, which can be easily discretized by choosing N points of the interval.

To minimize the error in the Euler equations, Telmer (1993) uses a Gauss–Seidel method. Recall that every Euler equation concerns two subsequent time periods and therefore portfolio terms for three time periods. For the third of these portfolios ("tomorrow's") and the corresponding prices, some decision rule as a function of the second ("today's") portfolio is assumed. For a given first portfolio ("yesterday's"), the algorithm now minimizes the Euler equation error by choosing the second portfolio. This ("today's") portfolio affects both terms in the Euler equation. Today's portfolio is computed for every point in

the state space. The error-minimizing portfolios and prices at every point in the state space lead to new updated decision rules. These new rules are then used for tomorrow's decision rules in the next iteration. This iterative process continues until the difference between two subsequent iterations (in an appropriate norm) is tiny. Actually, the difference in the portfolio functions will be exactly zero because of the discrete-valued nature of the functionals.

For this case of only one bond, Telmer can discretize the endogenous state space into 150 points per interval length of 0.1; he uses a total of more than 1,000 points. Because he considers only three exogenous shocks, the problem remains feasible. With such a fine discretization, the resulting errors turn out to be very low – for the interior of the interval of possible bond holdings, he reports maximum pricing errors of 0.0001 percent. Even though the errors are likely to be substantially higher at the boundary, for models with a one-dimensional state space, discrete methods generally achieve very good approximations. However, many interesting economic questions require a higher-dimensional state space.

9.1.2. Two Assets and the Curse of Dimensionality

Lucas (1994) and Heaton and Lucas (1996) use a discrete state space to approximate equilibria for the full model. They employ an auctioneer algorithm, which in essence is very similar to a Walrasian tatonnement process. Starting with an (approximating) policy function $\hat{f}_i$, the algorithm computes the supporting asset prices for both agents (which will be different). For that agent whose supporting stock price is higher, the stock holding is increased; for that agent whose supporting bond price is higher, the bond holding is increased. The amended portfolio holdings are then used for the new policy function $\hat{f}_{i+1}$. Iteration continues until the difference between the implied prices becomes sufficiently small, or cannot be improved further.

In principle, by employing a fine enough discretization, they could achieve a similar precision as Telmer in his simple setting. However, despite the speed of modern computers, it is still not nearly feasible to allow enough discrete points for the endogenous state space. Heaton and Lucas (1996) allow for 900 different holdings with 30 possible values in each dimension. This coarse discretization results in high errors. They report average (not maximum) errors of up to 0.4 percent. For the purposes of many economic insights, this might well be sufficient. In particular, the purpose of their paper is to investigate how missing asset markets might help to explain the equity premium puzzle, that is, the first moment of asset returns. It is very unlikely that the true equilibrium is so far away from their approximation that it has an influence on average returns (in fact, Judd et al., 1999a, repeat their calculations with a different algorithm and come to the same qualitative conclusions).

However, for many other applications that investigate welfare effects or higher moments of security prices, discrete methods are of limited use. It seems that running times would increase drastically if one tried to reduce these errors. The main problem lies in the fact that, for a discretized state space, the system

of Euler equations cannot be solved by efficient algorithms, such as Newton's method, which are designed for smooth systems. Instead they must be solved by some search procedure such as in Lucas' and Telmer's papers.

9.2. A Parameterized Expectations Algorithm Approach

The spline approximation method was a continuous state-space approach to the problem. This is a natural specification, because portfolios do not naturally fall on a finite grid of points and discretization methods suffer from the need to make the discretization small in order for the approximation to be acceptable. Marcet and Singleton (1999) also took a continuous state space approach.[6] They used Marcet's version of the parameterized expectations algorithm to solve a model with equity, a bond, and idiosyncratic shocks to income. Let W_i be the wealth of type i agents, p the exdividend price of equity, d the dividend, and y the current state of aggregate and idiosyncratic shocks. They focus on the expectations functions

$$\varphi_i(W_{1.t}, W_{2.t}, y_t) = E\{u_i'(c_{i,t+1})|W_{1,t}, W_{2.t}, y_t\}, \quad i = 1, 2,$$
$$\varphi_{2+i}(W_{1.t}, W_{2.t}, y_t) = E\{u_i'(c_{i.t+1})(p_{t+1} + d_{t+1})|W_{1,t}, W_{2.t}, y_t\}, \quad i = 1, 2$$

for next period's marginal utility of consumption and stocks. They use the Euler equations for bond and equity investment to fix consumption and asset holdings at time t in terms of the $\varphi(W_1, W_2, z)$ functions. This parameterization has the advantage that conditional expectations functions will be smoother than consumption and pricing functions because borrowing constraints may produce kinks in the relationship between consumption and wealth. This smoothing property of conditional expectations was first exploited by Wright and Williams (1982a, 1982b, 1984), for whom the nonnegativity constraint on commodity stockholding created similar kinks in consumption and price functions. Because borrowing constraints are similar to stockholding constraints, it is natural to also use the idea of Wright and Williams to parameterize a conditional expectation in this asset market context.

Marcet and Singleton approximate the φ functions with low-order exponential polynomials of wealth and the exogenous shocks of the form

$$\psi_j(\beta, W_1, W_2, y) = \exp(\beta_{j,1} + (\beta_{j,2}, \beta_{j,3}, \beta_{j,4}) \log y + \beta_{j,5} W_1 + \beta_{j,6} W_2). \tag{9.1}$$

They combined simulation methods and a successive approximation method to fix the β coefficients in (9.1). More precisely, they make a guess for the unknown

[6] It should be noted that Marcet and Singleton (1999) is a macroeconomic dynamics vintage article that is essentially the same as the 1991 working paper version, which predates most of the literature cited here.

coefficients; they simulate the process with those parameters. They use the data generated by the simulation to estimate the Euler equation errors implied by the candidate parameterization, and then they adjust the parameterization using the learning ideas of Marcet and Sargent (1989). This is repeated until the parameterization has converged.

Marcet and Singleton report that their algorithm had problems converging. This is not surprising. First, Figures 7.1 and 7.2 show that equilibrium behavior is not going to be well approximated by low-order polynomials. Second, the Euler equations for this problem tend to be ill conditioning because the assets are close substitutes. The ill conditioning implies that small changes in the coefficients will produce large changes in portfolio decisions and the Euler equation errors, making it difficult for any nonlinear equation procedure to converge, particularly the first order, successive approximation scheme used in Marcet and Singleton.

10. DYNAMIC MODELS WITH STRATEGIC POWER

Previous sections discuss only problems of competitive markets. There are many dynamic problems in which some agents have market power. Space limitations prevent a detailed presentation, but there are some papers that should be mentioned and that nicely illustrate the critical numerical techniques for dynamic games and illustrate the progress made in that literature.

A Markov perfect equilibrium (which was known as a closed-loop perfect state observation equilibrium in the older dynamic game literature) of a dynamic game is a solution to a system of functional equations similar to those in dynamic programming. Euler equations do not suffice because each player cares about the strategic reactions of the other players. One solves dynamic games by solving for each player's value function, where the value functions must satisfy a system of coupled Bellman equations; that is, each player has a Bellman equation modeling his or her decisionmaking process and takes as given the strategy of the other players. Equilibrium occurs when each player follows the strategy the other players expect him or her to follow. See Basar and Olsder (1995) for a recent presentation of dynamic games. Outside of some linear-quadratic cases, these models typically do not have closed-form solutions. Numerical methods are therefore crucial to analyzing these problems.

There were some efforts to solve strategic problems in the 1980s. An early example was the analysis by Wright and Williams (1982a) of the oil stockpiling policy of a government that knew that it would impose price controls if prices rose too high. Another example was the analysis by Kotlikoff, Shoven, and Spivak (1986) of strategic bequests. Both of these papers used polynomial approximation methods to solve the dynamic consistency problems. For example, Kotlikoff et al. solved a dynamic bargaining problem by approximating the value function with low-order polynomials. The approximation was

generated iteratively by guessing a value function, then computing, by means of a Bellman-like equation, the new guess for the value function at eighty points in the state space, and then using ordinary regression to generate the new guess for the players' value functions. They allowed their iterative schemes to continue until economic variables agreed in the first two significant digits. Because the scheme is at best linearly convergent, this indicates that the accuracy was somewhat less than two digits.

Although the efforts by Wright and Williams and Kotlikoff et al. were successful, the methods were slow and produced only moderately accurate answers. The more formal approach suggested by projection methods has recently been used to solve dynamic strategic problems more efficiently. In particular, projection methods have recently been used to compute time-consistent equilibria of government policy. Rui and Miranda (1996) use a projection method to solve a game between countries manipulating commodity storage policies to affect prices. Ha and Sibert (1997) used projection methods to solve games of tax policy competition between open economies. Both Rui and Miranda and Ha and Sibert used projection methods with orthogonal polynomials, and they had little difficulty in constructing stable and reliable algorithms. Vedenov and Miranda (2001) use collocation methods to solve dynamic duopoly models. The Euler and Bellman equation errors of their solutions were very small, of the order of 1 part in 10^5, indicating high accuracy in their solutions.

The success of these three recent papers indicates that much more complex dynamic games can be reliably and quickly solved numerically. They also emphasize our point that standard methods from numerical analysis can be used to create effective algorithms to solve complex economic models.

11. ASYMPTOTIC METHODS

Simple linearization methods are frequently used in dynamic economic analysis. More recently, more advanced versions have been introduced into the economic literature, producing more accurate and reliable approximations. They are similar to projection methods in that they try to compute a polynomial or similar approximation, but they use different information. Perturbation and asymptotic methods are constructive implementations of implicit function theorems. They are quite different from projection methods in terms of the underlying mathematics and the computer software needed to use them. We examine both regular perturbation methods for rational expectations models and an example of a problem in which bifurcation methods can solve a problem with a singularity.

The mathematical foundations for asymptotic methods in continuous time are presented in Fleming (1971) and Bensoussan (1988) and are applied to economic models in Magill (1977). This work has also proved useful in indicating how to apply perturbation methods to discrete-time models. We focus on the discrete-time methods here, but the same ideas apply to continuous-time models.

11.1. Regular Perturbations and Rational Expectations Models

We illustrate regular perturbation in the context of the basic rational expectations equations in a simple optimal growth model and then in the more general context of (3.1). Consider first the simple stochastic optimal growth problems

$$\max_{c_t} \quad \sum_{t=0}^{\infty} \beta^t u(c_t) \tag{11.1}$$
$$\text{s.t.} \quad k_{t+1} = F(k_t - c_t)(1 + \epsilon z_t),$$

where the z_t are iid with unit variance, and ϵ is a parameter expressing the standard deviation. The solution of the deterministic case, $\epsilon = 0$, can be expressed as a policy function, $C(k)$, satisfying the Euler equation

$$u'(C(k)) = \beta u'(C(F(k - C(k))))F'(k - C(k)).$$

Standard linearization methods (such as those described in Kydland and Prescott, 1982, and Laitner, 1981, 1984) produce $C'(k)$. Judd and Guu (1993, 1997) show that we can do much better than the linear approximation if we compute more terms in the Taylor series expansion

$$C(k) = C(k^*) + C'(k^*)(k - k^*) + C''(k^*)(k - k^*)^2/2 + \cdots$$

Successive differentiations of (11.1) produce higher-order derivatives of $C(k)$ at $k = k^*$. For example, the second derivative of (11.1), together with the steady-state condition $k = k^*$, implies that $C''(k^*)$ satisfies the linear equation

$$\begin{aligned} u''C'' + u'''C'C' = {} & \beta u'''(C'F'(1 - C'))^2 F' + \beta u''C''(F'(1 - C'))^2 F' \\ & + 2\beta u''C'F'(1 - C')^2 F'' + \beta u'F'''(1 - C')^2 \\ & + \beta u'F''(-C''), \end{aligned}$$

where all functions are evaluated at the steady-state value of their arguments. Linear operations combined with successive differentiations of (11.1) produce all higher-order derivatives.

The solution in the general stochastic case is a policy function, $C(k, \epsilon)$, which expresses consumption as a function of the capital stock k as well as the standard deviation ϵ. $C(k, \epsilon)$ satisfies the Euler equation

$$u'(C(k)) = \beta E\{u'(g(\epsilon, k, z))\, R(\epsilon, k, z)\}, \tag{11.2}$$

where

$$g(\epsilon, k, z) \equiv C((1 + \epsilon z)F(k - C(k))),$$
$$R(\epsilon, k, z) \equiv (1 + \epsilon z)F'(k - C(k)).$$

Differentiation of (11.2) shows that

$$C_\epsilon = 0,$$
$$C_{\epsilon\epsilon} = \frac{u'''C'C'F^2 + 2u''C'F + u''C''F^2}{u''C'F' + \beta u'F''},$$

where all the derivatives of u and F are evaluated at the steady-state values of c and k. This can be continued to compute higher-order derivatives as long as u and F have the necessary derivatives.

Judd and Guu (1993, 1997) apply these ideas to single-state deterministic and stochastic problems in continuous time and deterministic problems in discrete time. They show that the Euler equation errors are reduced by many orders of magnitude compared with the standard linear approximation, and that higher-order Taylor series are good approximations over a much wider range of state variables than the linear approximation.

This approach can also be used to analyze multidimensional problems. Judd and Gaspar (1997) examine the continuous-time case, and Judd (1998) presents an analysis of discrete-time problems. In general, if there are several endogenous variables, $Y(x)$, which are functions of a multidimensional state variable x, then we can compute the steady state, compute the linearization $Y(x)$ through standard eigenvalue decomposition methods from linear rational expectations, and then proceed as above to compute the higher-order derivatives of Y. To use a perturbation method, we express (3.1) in the form

$$\begin{aligned} 0 &= E\{g(x_t, y_t, x_{t+1}, y_{t+1}, \epsilon) \mid x_t\}, \\ x_{t+1} &= F(x_t, y_t, \epsilon z_t), \end{aligned} \tag{11.3}$$

where ϵ is a scaling parameter for the disturbance terms z. If the components of z_t have unit variance, then ϵ is the standard deviation. Different values for ϵ represent economies with different disturbances. The key observation is that we often know much about the $\epsilon = 0$ economy because (11.3) reduces to a deterministic problem. We build on that fact by using implicit function theorems.

The objective is to find some equilibrium rule, $\hat{Y}(x, \epsilon)$, such that

$$E\{g(x, \hat{Y}(x, \epsilon), F(x, \hat{Y}(x, \epsilon), \epsilon z), \hat{Y}(F(x, \hat{Y}(x, \epsilon), \epsilon z), \epsilon) \mid x\} \doteq 0.$$

Perturbation methods aim to approximate $Y(x, \epsilon)$ with a polynomial, just as projection methods do.[7] However, perturbation methods fix the unknown coefficients by computing the derivatives of $Y(x, \epsilon)$ at some value of (x, ϵ) where we know exactly $Y(x, \epsilon)$. Perturbation methods begin with the deterministic steady state, which is the solution to

$$\begin{aligned} &g(x^*, y^*, x^*, y^*, 0) = 0, \\ &x^* = F(x^*, y^*, 0). \end{aligned}$$

The objective is to find the derivatives of $Y(x, \epsilon)$ with respect to x and ϵ at the deterministic steady state, and use that information to construct Taylor series

[7] Perturbation methods can be used more generally to construct approximations that are nonlinear in their coefficients, but this is seldom done in economics. See Judd and Guu (1997) for an example of where Pade approximations are generated from perturbation data and significantly outperform standard Taylor series expansions.

approximations of $Y(x, \epsilon)$, such as

$$Y(x, \epsilon) \doteq y^* + Y_x(x^*, 0)(x - x^*) + Y_\epsilon(x^*, 0)\epsilon + (x - x^*)' Y_{xx}(x^*, 0)(x - x^*) + \cdots.$$

The second step in perturbation methods is to compute the linear terms of the approximation $Y_x(x^*, 0)$. Standard linearization methods show that the coefficients $Y_x(x^*, 0)$ are the solution, $y = Y_x x$, to the linear rational expectations model

$$g_1 x_t + g_2 y_t + g_3 x_{t+1} + g_4 y_{t+1} = 0, \tag{11.4}$$

where all the gradients of g in (11.4) are evaluated at the deterministic steady state. See Blanchard and Kahn (1980) and the survey by Anderson, Hansen, and McGrattean (1996) on methods for solving such models. This is the difficult step computationally, but it can be handled by conventional methods.

Computing the higher-order derivatives, such as $Y_{xx}(x^*, 0)$, $Y_\epsilon(x^*, 0)$, $Y_{x\epsilon}(x^*, 0)$, and so on, is actually easier than computing $Y_x(x^*, 0)$ because they are solutions to linear algebraic equations. Judd and Guu (1993) show how to use perturbation methods to solve simple one-sector optimal growth problems, and they show that the results are very accurate in that the Euler equation errors are small. Judd and Gaspar (1997) exposit the details of applying perturbation methods to optimal control problems.

Perturbation methods are the only methods that can handle problems with large dimension. Projection methods will suffer from various curses of dimensionality, because the number of unknown coefficients in the policy and value functions becomes large, the number of nonlinear equations used to identify the unknown coefficients becomes large, and the conditional expectations are multidimensional integrals. The combination of those factors makes it difficult to solve multiagent problems similar to (3.4) for more than a few agents. Judd and Gaspar find that problems with five agents and one asset are tractable, but their computations indicate that this is close to the limit of technology at that time. Perturbation methods have the advantage of reducing the integral calculations to moments of the random variables and linear operations. Therefore, large problems become more feasible.

11.2. Bifurcation Methods for Small Noise Portfolio Problems

The problems just discussed began with a locally unique steady state for the deterministic version. Many interesting problems with heterogeneous agents lack a unique steady state, implying that the techniques discussed herein do not apply directly. One such case is that in which there are multiple assets traded among agents with different tastes for risk. Suppose, for example, that there is trade in both a risky equity asset and a safe bond. In the deterministic steady state, all assets must have the same returns, implying that investors are indifferent among various assets. Hence, there is not a unique steady-state

holding of assets, even though the equilibrium holding of assets may be unique whenever there are positive amounts of risk.

There have been some attempts to use simple linear quadratic approximation methods such as the perturbation methods described herein. For example, Tesar (1995) used a linear-quadratic approach to evaluate the utility impact on countries of opening up trade in a bond. Each country had a stochastic endowment and the issue was how much risk sharing could be accomplished only through trade in a bond. This was a model with only one asset. One of her examples indicated that moving from trading in one bond, a case of incomplete asset markets, to complete markets would result in a Pareto inferior allocation, a finding that contradicts the first welfare theorem of general equilibrium. This example illustrates the need for using methods from the mathematical literature instead of relying on ad hoc approximation procedures. Again, we see the value of Frisch's observation that mathematics is necessary for safe and consistent analyses.

More typically we would like to solve models with multiple assets. Judd (1996, 1998) and Judd and Guu (2001) describe bifurcation methods, and they show that the basic portfolio demand problem is a good example where bifurcation methods can be used. Suppose that an investor has W in wealth to invest in two assets. The safe asset yields R per dollar invested and the risky asset yields Z per dollar invested. If a proportion ω of the investor's wealth is invested in the risky asset, final wealth is $Y = W((1-\omega)R + \omega Z)$. We assume that he or she chooses ω to maximize $E\{u(Y)\}$ for some concave, von Neumann–Morgenstern utility function $u(\cdot)$.

We want to "linearize" around the deterministic case. To do this, we parameterize the problem in terms of a scaling parameter ϵ and compute a Taylor series expansion for asset demand around the case of $\epsilon = 0$. The first problem we encounter is that if we eliminate risk by replacing Z with its mean, $\bar{Z}$, the resulting problem is unbounded if $R \neq \bar{Z}$ and indeterminate if $R = \bar{Z}$. Because the former case is untenable, we opt for the latter. We create a continuum of portfolio problems by assuming

$$Z = R + \epsilon z + \epsilon^2 \pi, \tag{11.5}$$

where $E\{z\} = 0$. At $\epsilon = 0$, Z is degenerate and equal to R. We assume that $\pi > 0$ because risky assets pay a premium. Note that we multiply z by ϵ and π by ϵ^2. Because the variance of ϵz is $\epsilon^2 \sigma_z^2$, this models the standard result in finance that risk premia are roughly proportional to variance.

We now investigate the collection of portfolio problems indexed by ϵ in (11.5). The first-order condition for ω, after dividing by ϵW, is, for all ϵ, equivalent to

$$0 = E\{u'(WR + \omega W(\epsilon z + \epsilon^2 \pi))(z + \epsilon \pi)\} \equiv G(\omega, \epsilon). \tag{11.6}$$

Equation (11.6) defines the solution to the asset demand problem even when $\epsilon = 0$. We know from concavity of $u(c)$ that there is a unique solution to (11.6) for ω if $\epsilon \neq 0$. However, at $\epsilon = 0$, ω can be anything because the two assets are

perfect substitutes. The indeterminacy of ω at $\epsilon = 0$ follows from the fact that $0 = G(\omega, 0)$ for all ω.

We want to solve for $\omega(\epsilon)$ as a Taylor series in ϵ. If there is such a series, implicit differentiation implies

$$0 = G_\omega \omega' + G_\epsilon, \tag{11.7}$$

where

$$G_\epsilon = E\{u''(Y)W(\omega z + 2\omega\epsilon\pi)W(z + \epsilon\pi) + u'(Y)\pi\},$$
$$G_\omega = E\{u''(Y)W(z + \epsilon\pi)^2\epsilon\}.$$

At $\epsilon = 0$, $G_\omega = 0$ for all ω. This implies that at no point $(\omega, 0)$ can we apply the implicit function theorem to (11.7) to solve for $\omega'(0)$. Moreover, we do not know $\lim_{\epsilon\to 0} \omega(\epsilon)$. However, let us proceed as if we can apply the implicit function theorem. Then (11.7) can be written as

$$\omega' = -\frac{G_\epsilon}{G_\omega}.$$

This looks bad because $G_\omega = 0$ at $\epsilon = 0$ until we remember l'Hôpital's rule. Suppose that we found a point ω_0 satisfying

$$0 = G_\epsilon(\omega_0, 0).$$

Then l'Hôpital's rule says that

$$\omega' = -\frac{G_{\epsilon\epsilon}}{G_{\omega\epsilon}},$$

which is well defined as long as $G_{\omega\epsilon} \neq 0$. So, let us proceed in this way. At $\epsilon = 0$, the second derivative of (11.6) with respect to ϵ reduces to $0 = u''(RW)\omega_0\sigma_z^2 W + u'(RW)\pi$, which implies that

$$\omega_0 W = -\frac{u'(WR)}{u''(WR)}\frac{\pi}{\sigma_z^2}. \tag{11.8}$$

Formula (11.8) is the simple portfolio rule from linear-quadratic analysis, saying that the total demand for the risky asset, $\omega_0 W$, is the product of risk tolerance and the risk premium per unit variance. However, we must check that it is consistent with the original problem and its first-order condition (11.6). For example, suppose that the support of z were unbounded above (as with a log-normal distribution) and that ω_0 exceeds one (which will happen if the Sharpe ratio and/or risk tolerance were sufficiently large). Then the support of $Y = W((1 - \omega_0)R + \omega_0 Z)$ would include negative values. But, if $u(c)$ were CRRA, then negative values of Y would never be chosen. Samuelson (1970) pointed out this problem and argued that we should restrict our attention to z with bounded support. Judd and Guu (2001) put this problem in the context of an implicit function theorem and derive analyticity conditions that validate this procedure. In any case, one must check that ω_0 is a choice consistent with (11.6) before we can continue.

Even if it is a valid solution, ω_0 is not an approximation to the portfolio choice at any particular variance. Instead, ω_0 is the limiting portfolio share as the variance vanishes. If we want the linear and quadratic approximations of $\omega(\epsilon)$ at $\epsilon = 0$, we must go further, because the quadratic approximation to $\omega(\epsilon)$ is $\omega(\epsilon) \doteq \omega(0) + \epsilon\omega'(0) + (\epsilon^2/2)\omega''(0)$. To calculate $\omega'(0)$ and $\omega''(0)$, we need to do two more rounds of implicit differentiation with respect to ϵ. If we differentiate (11.6) twice with respect to ϵ, we find that

$$0 = G_{\omega\omega}\omega'\omega' + 2G_{\omega\epsilon}\omega' + G_\omega\omega'' + G_{\epsilon\epsilon},$$

where (without loss of generality, we assume that $W = 1$)

$$\begin{aligned}
G_{\epsilon\epsilon} &= E\{u'''(Y)(\omega z + 2\omega\epsilon\pi)^2(z + \epsilon\pi) + u''(Y)2\omega\pi(z + \epsilon\pi) \\
&\quad + 2u''(Y)(\omega z + 2\omega\epsilon\pi)\pi\}, \\
G_{\omega\omega} &= E\{u'''(Y)(z + \epsilon\pi)^3\epsilon\}, \\
G_{\omega\epsilon} &= E\{u'''(Y)(\omega z + 2\omega\epsilon\pi)(z + \epsilon\pi)^2\epsilon + u''(Y)(z + \epsilon\pi)2\pi\epsilon \\
&\quad + u''(Y)(z + \epsilon\pi)^2\}.
\end{aligned}$$

At $\epsilon = 0$, $G_{\epsilon\epsilon} = u'''(R)\omega_0^2 E\{z^3\}$, $G_{\omega\omega} = 0$, and $G_{\omega\epsilon} = u''(R)E\{z^2\} \neq 0$. Therefore, l'Hôpital's rule applies and

$$\omega' = -\frac{1}{2}\frac{u'''(R)}{u''(R)}\frac{E\{z^3\}}{E\{z^2\}}\omega_0^2. \tag{11.9}$$

Equation (11.9) is a simple formula. It shows that, as riskiness increases, the change in ω depends on u'''/u'' and the ratio of skewness to variance. If u is quadratic or z is symmetric, ω does not change to a first order. We could continue this and compute more derivatives of $\omega(\epsilon)$ as long as u is sufficiently differentiable.

We end the development of this example here. However, it is clear that much more can be done. Judd and Guu (2001) develop this approach by using formal tools from bifurcation theory and apply it to problems of asset demand with several assets, asset equilibrium with incomplete asset markets, and problems of optimal asset innovation. Because the key bifurcation theorems are also true in Banach spaces, these methods can presumably be used to approximate equilibria in stationary dynamic models.

12. CONCLUSION

The work completed over the past decade indicates that there is steady progress in computing equilibria in markets with several agents. The key tools have been exploitation of tools from numerical analysis such as approximation theory, quadrature methods, and methods for solving large systems of equations. These efforts now make it tractable to accurately solve many dynamic markets with complete or incomplete asset markets, and even problems with strategic interactions, such as solving for time-consistent policies.

Further progress is likely because economists have just begun to exploit the full range of available numerical tools. For example, economists are just beginning to make use of methods, such as perturbation and asymptotic methods, that combine symbolic and numerical methods. Also, more advanced numerical and symbolic methods make higher-dimensional problems more tractable. The combination of faster computers, more efficient algorithms, and user-friendly software development tools means that this area of economic research will continue to grow and become an increasingly useful tool in analyzing complex dynamic economic models.

References

Anderson, E., L. P. Hansen, and E. McGrattan (1996), "Mechanics of Forming and Estimating Dynamic Linear Economies," in *Handbook of Computational Economics: Volume I*, (ed. by H. M. Amman, D. A. Kendrick, and J. Rust), Amsterdam: Elsevier Science, North-Holland.

Basar, T. and G. L. Olsder (1995), *Dynamic Non-Cooperative Game Theory*, 2nd ed. New York: Academic Press.

Bensoussan, A. (1988), *Perturbation Methods in Optimal Control*. New York: Wiley.

Blanchard, O. J. and C. M. Kahn (1980), "The Solution of Linear Difference Models under Rational Expectations," *Econometrica*, 48, 1305–1311.

Blum, L., F. Cucker, M. Shub, and S. Smale (1998), *Complexity and Real Computation*. New York: Springer Verlag.

Boucekkine, R. (1995), "An Alternative Methodology for Solving Nonlinear Forward-Looking Models," *Journal of Economic Dynamics and Control*, 19, 711–734.

deBoor, C. (1978), *A Practical Guide to Splines*. New York: Springer Verlag.

den Haan, W. (1997), "Solving Dynamic Models with Aggregate Shocks and Heterogeneous Agents," *Macroeconomic Dynamics*, 1, 355–386.

Eaves, B. C. and Schmedders, K. (1999), "General Equilibrium Models and Homotopy Methods," *Journal of Economic Dynamics and Control*, 23, 1249–1279.

Fair, R. C. and J. B. Taylor (1983), "Solution and Maximum Likelihood Estimation of Dynamic Nonlinear Rational Expectations Models," *Econometrica*, 51, 1169–1185.

Fisher, P. G. and A. J. Hughes Hallett (1987), "The Convergence Characteristics of Iterative Techniques for Solving Econometric Models," *Oxford Bulletin of Economics and Statistics*, 49, 231–244.

Fisher, P. G. and A. J. Hughes Hallett (1988), "Iterative Techniques for Solving Simultaneous Equation Systems: A View From the Economics Literature," *Journal of Computational and Applied Mathematics*, 24, 241–255.

Fleming, W. (1971), "Stochastic Control for Small Noise Intensities," *SIAM Journal of Control*, 9, 473–517.

Garcia, C. and W. Zangwill (1981), *Pathways to Solutions, Fixed Points, and Equilibria*. Englewood Cliffs, NJ: Prentice-Hall.

Gilli, M. and G. Pauletto (1998), "Krylov Methods for Solving Models with Forward-Looking Variables," *Journal of Economic Dynamics and Control*, 22, 1275–1289.

Gustafson, R. L. (1958), "Carryover Levels for Grains: A Method for Determining Amounts That are Optimal under Specified Conditions," USDA Technical Bulletin 1178.

Ha, J. and A. Sibert (1997), "Strategic Capital Taxation in Large Open Economies with Mobile Capital," *International Tax and Public Finance*, 4, 243–262.

Heaton, J. and D. J. Lucas (1996), "Evaluating the Effects of Incomplete Markets on Risk Sharing and Asset Pricing," *Journal of Political Economy*, 104, 443–487.

Hughes Hallett, A. J. and L. Piscitelli (1998), "Simple Reordering Techniques for Expanding the Convergence Radius of First-Order Iterative Techniques," *Journal of Economic Dynamics and Control*, 22, 1319–1333.

Judd, K. L. (1992), "Projection Methods for Solving Aggregate Growth Models," *Journal of Economic Theory*, 58, 410–452.

Judd, K. L. (1996), "Approximation, Perturbation, and Projection Methods in Economic Analysis," in *Handbook of Computational Economics*, (ed. by H. Amman, D. Kendrick, and J. Rust), Amsterdam: North-Holland.

Judd, K. L. (1998), *Numerical Methods in Economics*. Cambridge, MA: MIT Press.

Judd, K. L. (2002), "Parametric Path Method for Solving Perfect Foresight Models," *Journal of Economic Dynamics and Control*, 26, 1557–1583.

Judd, K. L. and J. Gaspar (1997), "Perturbation Methods for Discrete-Time Dynamic Deterministic Models," *Macroeconomic Dynamics*, 1, 45–75.

Judd, K. L. and S.-M. Guu (1993), "Perturbation Solution Methods for Economic Growth Models," in *Economic and Financial Modeling With Mathematica*, (ed. by Hal Varian), New York: Springer-Verlag.

Judd, K. L. and S.-M. Guu (1997), "Asymptotic Methods for Aggregate Growth Models," *Journal of Economic Dynamics and Control*, 21, 1025–1042.

Judd, K. L. and S.-M. Guu (2001), "Asymptotic Methods for Asset Market Equilibrium Analysis," *Economic Theory*, 18, 127–157.

Judd, K. L., F. Kubler, and K. Schmedders (1999a), "The Impact of Portfolio Constraints in Infinite-Horizon Incomplete-Markets Models," in *The Theory of Markets*, (ed. by P. J. J. Herings, A. J. J. Talman, and G. van der Laan), Amsterdam: North-Holland.

Judd, K. L., F. Kubler, and K. Schmedders (1999b), "A Solution Method for Incomplete Asset Markets with Heterogeneous Agents," Working Paper, Hoover Institution, Stanford, CA.

Judd, K. L., F. Kubler, and K. Schmedders (2000), "Computing Equilibria in Infinite Horizon Finance Economies I: The Case of One Asset," *Journal of Economic Dynamics and Control*, 24, 1047–1078.

Juillard, M. (1996), "DYNARE: A Program for the Resolution and Simulation of Dynamic Models with Forward Variables through the Use of a Relaxation Algorithm," Working Paper 9602, CEPREMAP, Paris, France.

Juillard, M., D. Laxton, P. McAdam, and H. Pioro (1998), "An Algorithm Competition: First-Order Iterations Versus Newton-Based Techniques," *Journal of Economic Dynamics and Control*, 22, 1291–1318.

Kelley, C. T. (1995), *Iterative Methods for Linear and Nonlinear Equations*. Philadelphia: SIAM.

Kotlikoff, L., J. Shoven, and A. Spivak (1986), "The Effect of Annuity Insurance on Savings and Inequality," *Journal of Labor Economics*, 4, S183–S207.

Krusell, P. and A. A. Smith, Jr. (1997), "Income and Wealth Heterogeneity, Portfolio Choice, and Equilibrium Asset Returns," *Macroeconomic Dynamics*, 1, 387–422.

Kubler, F. and K. Schmedders, (2002), "Recursive Equilibria with Incomplete Financial Markets," *Macroeconomic Dynamics*, 6, 284–306.

Kydland, F. E. and E. C. Prescott (1982), "Time to Build and Aggregate Fluctuations," *Econometrica*, 50, 1345–1370.

Laffargue, J. P. (1990), "Resolution d'un Mode'le Macroeconomique avec Anticipations Rationnelles," *Annales d'Economie et Statistique*, 17, 97–119.

Laitner, J. (1981), "The Stability of Steady States in Perfect Foresight Models," *Econometrica*, 49, 319–333.

Laitner, J. (1984), "Transition Time Paths for Overlapping-Generations Models," *Journal of Economic Dynamics and Control*, 7, 111–129.

Levine, D. and W. Zame (1996), "Debt Constraint and Equilibrium in Infinite Horizon Economies with Incomplete Markets," *Journal of Mathematical Economics*, 26, 103–131.

Lucas, D. J. (1994), "Asset Pricing With Undiversifiable Income Risk and Short-Sale Constraints: Deepening the Equity Premium Puzzle," *Journal of Monetary Economics*, 34, 325–241.

Lucas, R. E. Jr. (1978), "Asset Prices in an Exchange Economy," *Econometrica*, 46, 1429–1445.

Magill, J. P. M. (1977), "A Local Analysis of *N*-Sector Capital Accumulation Under Uncertainty," *Journal of Economic Theory*, 15, 211–219.

Magill, J. P. M. and M. Quinzii (1996), "Incomplete Markets over an Infinite Horizon: Long-Lived Securities and Speculative Bubbles," *Journal of Mathematical Economics*, 26, 133–170.

Marcet, A. and T. Sargent (1989), "Convergence of Least Squares Learning Mechanisms in Self-Referential Linear Stochastic Models," *Journal of Economic Theory*, 48, 337–368.

Marcet, A. and K. Singleton (1999), "Equilibrium Asset Prices and Savings of Heterogeneous Agents in the Presence of Portfolio Constraints," *Macroeconomic Dynamics*, 3, 243–277.

Miranda, M. J. and P. G. Helmberger (1988), "The Effects of Commodity Price Stabilization Programs," *American Economic Review*, 78, 46–58.

Rios-Rull, V. (1999), "Computation of Equilibria in Heterogenous Agent Models," in *Computational Methods for the Study of Dynamic Economics: An Introduction*, (ed. by R. Marimon and A. Scott), Oxford: Oxford University Press.

Rui, X. and M. J. Miranda (1996), "Solving Nonlinear Dynamic Games via Orthogonal Collocation: An Application to Strategic Commodity Buffer Stock Policy," *Annals of Operations Research*, 68, 89–108.

Saad, Y. (1996), *Iterative Methods for Sparse Linear Systems*, Boston, MA: PWS Publishing.

Samuelson, P. A. (1970), "The Fundamental Approximation Theorem of Portfolio Analysis in Terms of Means, Variances, and Higher Moments," *Review of Economic Studies*, 37, 537–542.

Santos, M. S. and J. Vigo-Aguiar (1998), "Analysis of a Numerical Dynamic Programming Algorithm Applied to Economic Models," *Econometrica*, 66, 409–426.

Schmedders, K. (1998), "Computing Equilibria in the General Equilibrium Model with Incomplete Asset Markets," *Journal of Economic Dynamics and Control*, 22, 1375–1401.

Taylor, J. B. and H. Uhlig (1990), "Solving Nonlinear Stochastic Growth Models: A Comparison of Alternative Solution Methods," *Journal of Business and Economic Statistics*, 8, 1–18.

Telmer, C. I. (1993), "Asset-Pricing Puzzles in Incomplete Markets," *Journal of Finance*, 48, 1803–1832.

Tesar, L. L. (1995), *Evaluating the Gains from International Risksharing*, Vol. 42, Carnegie–Rochester Conference Series on Public Policy, North Holland, 95–143.

Vedenov, D. and M. J. Miranda (2001), "Numerical Solution of Dynamic Oligopoly Games With Capital Investment," *Economic Theory*, 18, 237–261.

Wright, B. D. and J. C. Williams (1982a), "The Roles of Public and Private Storage in Managing Oil Import Disruptions," *Bell Journal of Economics*, 13, 341–353.

Wright, B. D. and J. C. Williams (1982b), "The Economic Role of Commodity Storage," *Economic Journal*, 92, 596–614.

Wright, B. D. and J. C. Williams (1984), "The Welfare Effects of the Introduction of Storage," *Quarterly Journal of Economics*, 99, 169–182.

Young, D. M. (1971), *Iterative Solution of Large Linear Systems*. New York: Academic Press.

Zhang, H. H. (1997a), "Endogenous Borrowing Constraint with Incomplete Markets," *Journal of Finance*, 52, 2187–2209.

Zhang, H. H. (1997b), "Endogenous Short-Sale Constraint, Stock Prices, and Output Cycles," *Macroeconomic Dynamics*, 1, 228–254.

Index