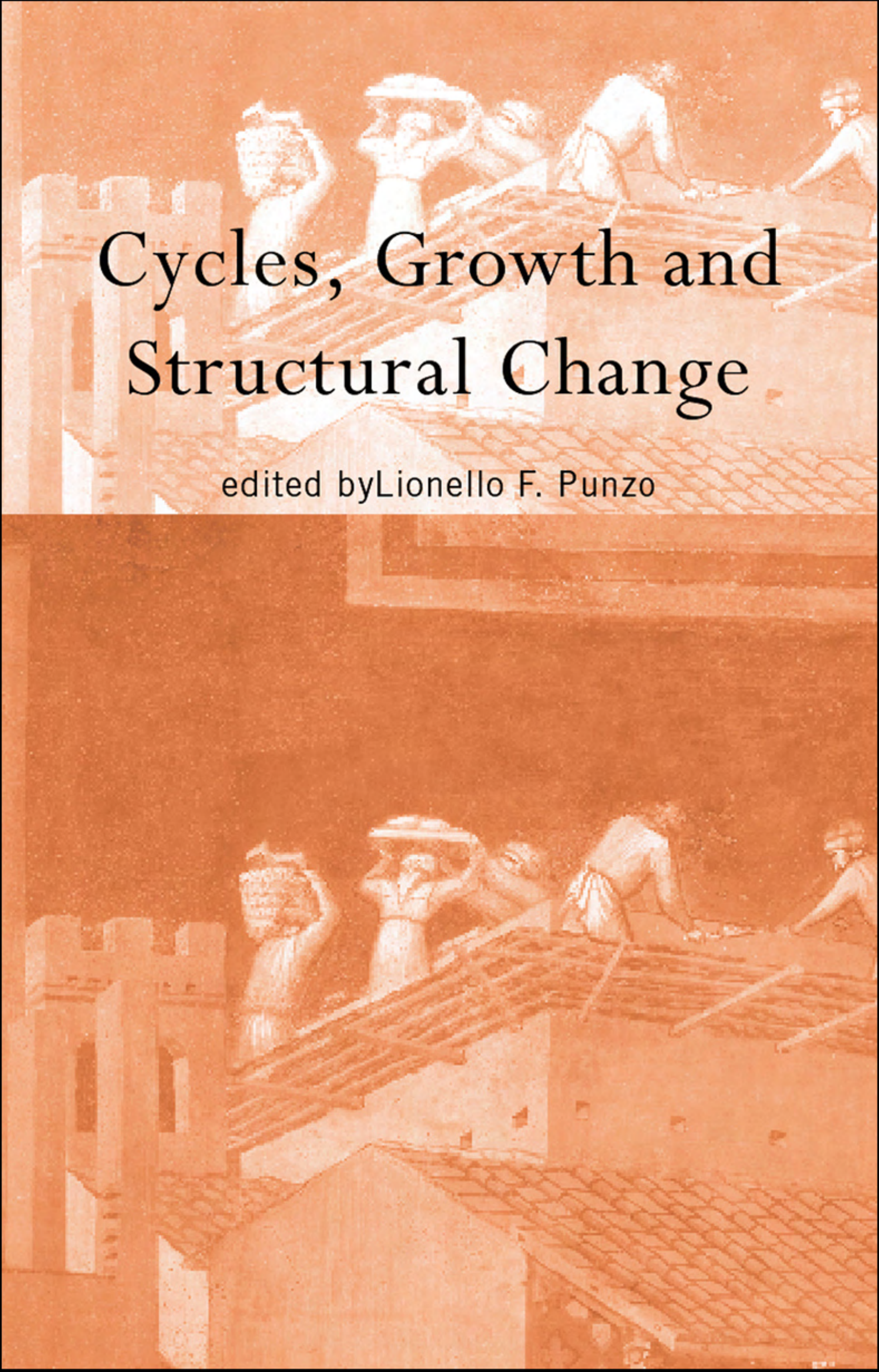


# Cycles, Growth and Structural Change

edited byLionello F. Punzo

# Cycles, Growth and Structural Change

With its unique combination of theoretical models and empirical analyses, this collection of specialists' essays contributes to an assessment of our understanding of the relationship between cycles, growth and structural change. In this light it also offers a critical evaluation of results in the recent literature on complex and chaotic dynamics. The book is constructed upon the hypothesis that a re-unification of the economic analyses of dynamic phenomena can be obtained through comparison and cross fertilization of their up-to-now specialized approaches.

Recent developments in economic dynamics, one of the exciting frontiers in current economic research, have been drawing from very diverse research areas and different specialisms: economic theorists, applied economists and econometricians, of course, but also mathematicians and researchers in information theory, time series and dynamical system theories. The variety of approaches in this book well represents the economists' views of this research area. They range from neoclassical and endogenous growth theories, to classical, evolutionary and neo-Austrian dynamics; from real business cycle and information-based econometrics to other non-linear techniques related to complex dynamics. Likewise, modelling is done in the macro-, micro-, multi-agent and large dynamical system styles. Moreover, the book offers materials for an evaluation of recent literature which relies on complexity properties to search for alternative formulations of economic policy issues.

Providing both historical evidence and theoretical formulations, written in a critical vein by leading specialists in their own areas, *Cycles, Growth and Structural Change* offers state-of-the art surveys of a fast expanding field and suggests a research agenda for the future.

**Lionello F. Punzo** is Professor of Economics at the University of Siena.

# Routledge Siena Studies in Political Economy

The Siena Summer School hosts lectures by distinguished scholars on topics characterized by a lively research activity. The lectures collected in this series offer a clear account of the alternative research paths that characterize a certain field. Former workshops of the School were printed by different publishers. They include:

**Macroeconomics**
A survey of research strategies
*Edited by Alessandro Vercelli and Nicola Dimitri*
Oxford University Press, 1992

**International Problems of Economic Interdependence**
*Edited by Massimo Di Matteo, Mario Baldassarri and Robert Mundell*
Macmillan, 1994

**Ethics, Rationality and Economic Behaviour**
*Edited by Francesco Farina, Frank Hahn and Stefano Vannucci*
Clarendon Press, 1996

**The Politics and Economics of Power**
*Edited by Samuel Bowles, Maurizio Franzini and Ugo Pagano*
Routledge, 1998

**The Evolution of Economic Diversity**
*Edited by Antonio Nicita and Ugo Pagano*
Routledge, 2000

**Cycles, Growth and Structural Change**
Theories and empirical evidence
*Edited by Lionello F. Punzo*
Routledge, 2001

The Routledge Siena Studies in Political Economy Series gives a comprehensive access to the publications of the School, which emphasizes the common methodology employed in organizing the different workshops.

# Cycles, Growth and Structural Change

## Theories and empirical evidence

**Edited by**
**Lionello F. Punzo**

London and New York

First published 2001
by Routledge
11 New Fetter Lane, London EC4P 4EE

Simultaneously published in the USA and Canada
by Routledge
29 West 35th Street, New York, NY 10001

*Routledge is an imprint of the Taylor & Francis Group*

This edition published in the Taylor & Francis e-Library, 2006.

*British Library Cataloguing in Publication Data*
A catalogue record for this book is available from the British Library

*Library of Congress Cataloging in Publication Data*
Punzo, Lionello F.
Cycles, growth and structural change : theories and empirical evidence / Lionello F. Punzo.
p. cm.
Includes bibliographical references and index.
1. Business cycles–Congresses. 2. Economic development–Congresses. 3.
Economics–Congresses. 4. Business cycles–Research–Methodology–Congresses. 5. Economic development–Research–Methodology–Congresses. 6. Economics–Methodology–Congresses. I. Title.

HB3711 .P93 2001
330–dc21
00–051777

ISBN 0-203-16483-0 Master e-book ISBN

ISBN 0-203-25909-2 (Adobe eReader Format)
ISBN 0–415–25137–0 (Print Edition)

# Contents

# Figures

# Tables

# Contributors

**Mario Amendola** is Professor of Economic Analysis at the University of Rome 'La Sapienza'. He is the author of various books on the theories of capital and technical progress, on dynamic models, on technological innovation and industrial change, and of papers published on important international journals, and also carries out an intense applied research activity. He has recently published (with J.-L. Gaffard) *Out of Equilibrium* (Clarendon Press, 1998).

**Masanao Aoki** is Professor Emeritus at UCLA, and Professor in the Department of Economics and Information, Gifu Shotoku Gakuen University, Japan. His main current fields of research are illustrated by his most recent books, both from Cambridge University Press: *New Approaches to Macroeconomic Modeling: Evolutionary Stochastic Dynamics, Multiple Equilibria and Externalities as Field Effects* (1996) and *Modeling Aggregate Behavior and Fluctuations in Economics: Stochastic View of Interacting Agents* (forthcoming, 2001)

**William J. Baumol** is Professor of Economics at New York University and Professor Emeritus and Senior Research Scholar at Princeton University. His main current fields of interest are innovation and growth, public policy related to monopoly power and international trade in the presence of scale economies. He is the author of numerous books and articles, including *Productivity and American Leadership: The Long View* (with S.A. Batey Blackman and E.N. Wolff), (MIT Press, 1989).

**Bernhard Böhm** is Professor for Economics and Applied Econometrics at the Institute of Econometrics, Operations Research, and Systems Theory of the University of Technology, Vienna, Austria. He also chairs a postgraduate programme in applied economics at the Academia Istropolitana Nova in Slovakia. He specializes in econometric model building, structural change, and environmental economics.

**William A. Brock** is Vilas Research Professor of Economics at the University of Wisconsin, Madison. His interests include general economic dynamics, business cycle and monetary theory, heterogeneous agent models, econometrics, optimal growth theory, and finance. He is author of numerous articles and several books.

**Richard H. Day** is Professor of Economics at the University of Southern California. His research began with simulations of production, investment and technological change in various agricultural regions and industrial sectors using a class of recursive programming models based on bounded rationality and adaptive economizing. Subsequent work has involved the theory of adaption and economic evolution, business cycles, economic growth and economic development in the very long run. He is author of *Complex Economic Dynamics vol 1: Introduction to Dynamical Systems and Market Mechanisms; vol 2: Introduction to Macroeconomic Dynamics* (MIT Press, 1994 and 2001).

**Domenico Delli Gatti** is Economics Professor, Catholic University, Milan, Italy. He is the associate editor of the JEBO and has published articles on financial fragility, monetary economics, and non-linear dynamics. He has published a book on *Money, accunulation and Cycles* (NIS, 1994) and more recently edited a book on *Interaction and Market Structure* (Springer, 2000).

**Peter Flaschel** is Professor of Economics, Bielefeld University and Visiting Professor of Economics, University of Technology, Sydney. His research interests include disequilibrium, cycles and growth, open economy, macrodynamics, microdynamics, macroeconometric model building, numerical analysis of high order macrosystems. He is the author of several published books and articles.

**Jean-Luc Gaffard** is Professor of Economics at the University of Nice-Sophia Antipolis and senior member of the Institut Universitaire de France. His major research interests are in Industrial Economics, Economics of Innovation, and Dynamic Economics. He has published many papers in scientific journals and several books including, *The Innovative Choice. An Economic Analysis of the Dynamics of Technology* (with Mario Amendola) (Blackwell, 1988), and *Out of Equilibrium* (with Mario Amendola) (Clarendon Press, 1998).

**Mauro Gallegati** is Professor of Economic Dynamics, University of Teramo, Italy. He has written extensively on financial fragility, business cycle fluctuations, and non-linear dynamics. He has published a book on *Fluctuations in Italy* (Giappichelli, 1998) and has edited two books, *Beyond the RA* (Elgar, 1999) and *Interaction and Market Structure* (Springer, 2000).

**Frank Hahn** is Professor Emeritus at the University of Cambridge, Professore Ordinario at the University of Siena and Fellow of Churchill College. He is the author of numerous papers and has published a number of books.

**Tad Hogg** is a member of the research staff at the Xerox Palo Alto Research Center. His current research includes developing institutional mechanisms promoting privacy and electronic commerce and investigating the dynamics of computational ecosystems consisting of processes interacting with an unpredictable environment.

**Bernardo A. Huberman** is the Scientific Director of the Hewlett Packard Sand Hill Laboratories, a research centre recently established by Hewlett Packard to explore a number of novel issues in the research of distributed systems and its organizational setting. He is co-winner of the 1990 CECOIA prize in Economics and Artificial Intelligence and he recently shared the IBM Prize of the Society for Computational Economics. Dr. Huberman is one of the creators of the field of ecology of computation, editor of a book on the subject, and an expert on the dynamics and growth of the internet.

**Katsuhito Iwai** is Professor of Economics at the University of Tokyo. His works include: *Disequilibrium Dynamics*, (Yale University Press, 1981), 'The Bootstrap Theory of Money,' *Structural Change & Economic Dynamics*, 7, (4), 1996, and 'Persons, Things and Corporations,' *American J. of Comparative Law*, 47, (4), 1999.

**Oleg V. Pavlov** is a postdoctoral fellow at the Boston University of School of Management. His research concentrated on complex economic dynamics and computer simulation in such fields as population economics, economic development and technological evolution. His current research and teaching focuses on economics of information technology.

**Edward C. Prescott** is Regents' Professor at the University of Minnesota and an advisor to the Federal Reserve Bank of Minneapolis. His current research is concerned with economic growth and development, which is synthesized and extended in Parente's and his monograph *Barriers to Riches* (MIT Press, 2000).

**Lionello F. Punzo** is Professor of Economics at Siena University. His current interests are in the theory and empirics of growth and complex dynamics. Among his works are *The Dynamics of a Capitalist Economy* (co-authored with Richard M. Goodwin) (Polity Press and Westview Press, 1987), *Economic Performance* (edited with B. Böhm), (Physika Verlag, 1992), *European Economies in Transition* (edited with O. Fabel and F. Farina), (Macmillan, 2000) and *Mexico Beyond NAFTA* (edited with M. Puchet), (Routledge, 2001).

**Jorma Rissanen** is a member of the research staff at IBM Almaden Research Center. He is Visiting Professor at Royal Holloway, University of London and Professor Emeritus in the Technical University of Tampere, Finland. His main research interest is in creating a theory of statistical modeling, as outlined in his contribution to this book.

**Solomos Solomou** is a Senior Lecturer in the Faculty of Economics at the University of Cambridge, and a Fellow of Peterhouse. His work includes *Phases of Economic Growth, 1850–1973: Kondratieff Waves and Kuznets Swings* (Cambridge University Press, 1987) and *Economic Cycles* (Manchester University Press, 1998).

**Hiroshi Yoshikawa** is Professor of Economics at the University of Tokyo. His research interests are in macroeconomics and the Japanese economy. He is the author of *Macroeconomics and the Japanese Economy* (Oxford University Press, 1995).

**Michael Youssefmir** Michael Youssefmir is Director of Product Management for ArrayComm, Inc's wireless Internet product line. Previously he worked in the Internet Ecologies Group at the Xerox Palo Alto Research Center. He holds a number of patents in the application of adaptive array technology to wireless systems.

# Preface

This volume presents contributions revolving around the general theme of the relationship, empirical as well as theoretical, between economic oscillations, growth and structural change. It is constructed upon a twofold hypothesis. A reunification of the analyses of the three phenomena may be obtained through a comparison and cross fertilisation of the existing specialised approaches. Moreover, the accomplishment of such a project is perhaps within reach, now. This preface will discuss such a hypothesis, outlying the structure of the book and the string connecting individual contributions.

Most chapters are revised versions of papers read at the International Summer School of the same title, held in Siena in July 1998. Some chapters have been added with the intent of broadening the view without however any attempt to reach exhaustiveness. The field, in fact, represents one of the most exciting frontiers in current economic research, calling for the individual as well as, more often, joint efforts of specialists from very diverse areas: economic theorists, real life applied economists, econometricians, of course, but also mathematicians and people working in information-theory related areas. Cross fertilisation is at one of its historical peaks, and it is taking place here more than elsewhere in the economic discipline. The School was concluded by a lively workshop dedicated to the works of the late Richard M. Goodwin, whose lifetime research has been dedicated to the themes of this book.

## Convergence: the new impetus to growth theory and growth empirics

Among the factors that brought about the present rich and stimulating, though still fluid, state of affairs, one should count the new impetus to growth theory and growth empirics coming from the debate on ‘convergence’ (further supported by recent quantitative advances in cross-national comparative databases). Indeed, growth has been traditionally an area where a division of labour between theoretical and applied economics (and economists) could not be enforced. Similarly, the actual experience of fluctuations along with growth was the background for the development of the classical macrodynamics (the set of theories of business fluctuations of the 1930s to the 1950s) and more recently of new classical macrodynamics. So, the intersection between theoretical work and empirical evidence has been remarkably extensive for fluctuations as well as for growth. Perhaps it has been greater than in other parts of economics, and in a sense this very aspect makes these theories so appealing to academic economists, but more so to policy makers and the general public alike. The questions raised by the opening of the new millennium, in a scenario of uncertainties about the future developments of the world as well as the national economies, can only increase such general interest. Does economics now offer answers to such grand questions as: how to support long run, smooth (or relatively smooth) as well as even growth? This is an open question, at present. The recent developments expose a varied scenario, where high instability is coupled with equally highly, cross-country, diversified growth performance.

Thus, there is an issue of growth and convergence, but we are also still looking for a satisfactory answer to the age-old issue of explaining why growth seems to marry with fluctuations. This is the coupling of trend and oscillation that focussed the reflection of Schumpeter and inspired the more mathematically inclined thinkers of classical macrodynamics. This is one of the issues addressed in the chapters of this book. It is done with diverse approaches. They range from neo-Classical, to neo-Austrian, to Schumpeterian and Evolutionist, and they vary between macro and micro. The book shows as well a variety of styles or methodologies: from historical reconstruction to quantitative studies (cluster analysis-type of approach to real business cycle econometrics to information-theoretic model theory), to the modelling of large artificial economies (often of heterogeneous interacting agents/firms/sectors), and qualitative econometrics. Such twofold variety reflects the philosophy of the series of International Summer Schools at Siena since their operation.

The list of factors for the present situation is far from short. It should include the unending dispute between exogenous and endogenous explanations of long run growth and fluctuations; the discovery of the possibility, besides multiple equilibria, of endogenous fluctuations and of chaotic, or non-periodic, behaviours even in simplest theoretical models, constructed upon the conventional equilibrium ingredients; finally, the formalization of a non-mainstream tradition, along Evolutionary, neo-Schumpeterian, or else neo-Austrian lines, to name a few.

I believe all this can be traced back to the key issue anticipated above: the need for a coordination of the analysis and explanation of short run economic variability with that of growth as a long run phenomenon and of these two with the analysis of structural change. It may be useful to spend a few words on how this issue presents itself now against the background of the history of its conceptual evolution, and to indicate the *Goodwin connection*.

## Cycles, growth and structural change

The three ‘wings’ of cycles, growth and structural change became part of one and the same conceptual design, in this century, synthesised by the Schumpeterian vision of a market economy growing only through fluctuations. In the so-called

Schumpeterian clock, economic time is irregularly struck by the swarms of innovations, i.e. clusters of technological and economic shocks erratically but systematically administered to the markets. This vision proposed the irregular oscillation as the unifying concept for dynamics but this got lost in the process of formalization by Hicks, Samuelson, Harrod, Tinbergen, Kalecki and others which created modern mathematical dynamics. (It only survived in Goodwin's lifetime work.) Then, attention shifted towards more orderly dynamics, the existence and stability of point-equilibria and of regular oscillations (limit cycles and the like). The eventual pre-eminence of this quest went along with the separation of the two issues, the explanation of growth from the explanation of the oscillations. In place of the single exogenously initialised mechanism of fluctuating growth, a dualism emerged between oscillations and growth, and in parallel the former was identified with the short run and the latter with the long run. The products of such a separation were the neo-classical theory of growth and the Hicks–Goodwin theory of self-sustained oscillations. Such a separation took also the shape of a methodological (or *weltanschauung's)* opposition of an exogenous versus an endogenous explanation of dynamics, the former associated with Solow's growth model, and the latter with the classical macrodynamics just mentioned (in particular, Goodwin's own version of it). There are many reasons for this divorce between two research lines that otherwise both descend from the same Harrod's model (Solow inserting a production function, Goodwin going non-linear). One is the shift in post-war theoretical interests towards equilibrium issues and hence the related stability. But, probably, a more profound conceptual reason lies in the historical failure of those who believed in the paradigm of the endogenously sustained oscillations, to accomplish the self-assigned task of building the grand theory of the trend *cum* oscillations. Perhaps, the task was simply overly ambitious for those times and the mathematics at hand. More realistically, neo-classical growth theory never played with such a 'dream', as acknowledged by Solow himself.

It is too well known how the exogenous theory of growth survived (though with alternating phases) while at the same time the endogenous theory of oscillations went out of fashion. From the 1960s till recently, the latter was superseded by a simpler account of oscillations, sometimes referred to as the linear econometric model. This rests upon the notion of a fundamentally stable equilibrium behavior for the economy subjected to exogenous and random disturbances. Thus, this new dynamic wisdom had both long and short run explained by fundamental forces that were not modelled by the economist himself: for growth they were determistic in nature, while they were stochastic for oscillations, so that they did not exclude each other and on the contrary could be combined through a sort of division of labour. Such a mechanical view of the working of an economy producing short lived oscillations as a result of passive behaviours on the part of economic agents was eventually questioned by both the first and the second monetarist waves, the latter producing the new classical theory of business fluctuations. But, although enriching the representation of decision making processes by taking into account expectation-induced adjustments, the dynamics of oscillations remained the response to exogenous shocks of an otherwise point-stable mechanism. In growth theory, on the other hand, Solow's paradigm was challenged by the blossoming of a body of models collectively going under the name of endogenous models. They are endogenous in the sense that they account for the possibility of a sustained long run growth on the basis of economic forces associated with the very basic mechanisms driving a market economy.

After the middle of the 1980s, (macro-)economists lived in a situation that can be depicted as the mirror image of the one at the beginning of the 1950s: growth theory had tendentially become endogenous, while business cycle theory (Monetarist version) was exogenous! Beyond this interchange of points of view, basically the two 'varieties of dynamics' were still being treated in different parts of economics; the explanation of one would not be the explanation of the other. The issue of conjugating trend and oscillation in a unified model was still there to be addressed. RBC theory, which bravely picked it up, can be seen as a provocatory but forceful attempt at realising such unification. Likewise, in the same light, one can appraise the birth and diffusion of Evolutionary and Schumpeterian modelling, neo-Austrian modelling, and some of the non-label literature on endogenous fluctuations and chaos that has emerged from the mid-1980s to the present day. One of the chapters in this volume goes over the issue to recall the RBC proposal of a (re-)unification of growth and fluctuations dynamics centred upon the equilibrium theory of growth in a stochastic environment. Other contributions present alternative viewpoints, with an out-of-equilibrium approach yielding naturally long run, persistent fluctuations in growth rates (as well as levels, if needed). The complex dynamics thus obtained does not distinguish between fluctuations and growth, the latter being but one simple type of the former. Their proposal of re-unification centred upon a generalised notion of growth cycle in a fundamentally deterministic framework can be contrasted to the essentially stochastic equilibrium approach of RBC. Likewise, the chapter on the positiveness of long-run profits in a evolutionary and Schumpeterian framework, shows such a possibility (excluded by Schumpeter himself) to be strongly related with the issue at stake.

In a sense, all the chapters in this book address the same question: how can we have both sustained growth and sustained oscillations? If the answer is positive, they should rest upon the same mechanism. Searching for the answer, some theories and models redefine the very notion of growth and, more often, of oscillations to accommodate frontier phenomena belonging to so-called complex dynamics. To see where the common mechanism could be, one ingredient is missing, though.

## Dynamics from structural change

In fact, parallel to the theoretical divorce of growth from fluctuations, there was a separation of formal dynamics from structural change. Taking the economic structure as given, dynamics became basically a theory of adjustment processes. Structural dynamics, on the other hand, was left out of a style of modelling too aggregative to afford a notion of structure richer than the description of some statistical regularity. Thus, the analysis of structural change migrated to development theory and it stayed there until recently. As remarked by some of the contributors to a previous volume in this series, structural change eventually has re-appeared under disguise in some of the models of the endogenous growth family where the line between monotonic growth and discontinuous development is never clearly marked. Even Lucas' critique can be read as a critique of a modelling approach that assumed structure and kept it outside theoretical consideration, treating it as exogenous in the diminutive sense that Hahn points out in this volume. As a result of the recent historical developments, structural change has re-appeared on the agenda of the dynamic theorist and empiricist of the First World economies, too.

The mechanism that may provide a unique explanation of self-sustained dynamics lies inside the structure of the economy. This hypothesis was present more than intuitively to the minds of the believers of the endogenous program when forging classical macrodynamics. This was naturally so. More recent theoretical developments have forced us to reconsider our understanding of what makes up an economic structure (to include expectations, decision rules, institutions and rules of game, etc.), but all the recent dynamic literature accepts the principle that it is the endogenous structure of the economy, a set of broadly defined interacting 'fundamentals' that determines the dynamics we observe. The increasing wealth of interesting results obtained in this area springs exactly from the re-consideration of the endogenous inner network that makes up an economy.

The re-consideration of the relationship between dynamics and structure follows different lines, some of which are represented in this volume. If we use modelling styles as classification criterion, we can focus upon the architecture of such models. In fact, if we think of models as computing devices, then their formal structures are architectures just like in hardware structures. Hence, we may compare models built upon a serial with models built upon a parallel architecture. In this light, the various lines can be grouped into only two views, reflecting distinct modelling strategies but also different images of the economic world. In one of them, the economy is seen as a large-scale replica of a 'standardized' economic agent, which is a complex entity by itself and therefore encapsulates whatever is complex in the economic process. Hence, either the agent is a metaphor for the system, or else, the latter is a metaphor of the former. This is a *serial view*, in that the model of the whole economy is a scaled-up copy of the model of the economic agent (or vice versa), interaction between agents adding but complications to the basic mechanism. The *alternative, parallel view* depicts the economic system as fundamentally a web of interacting parts (e.g. typically, agents but also institutions, markets, rules etc.). It is their interaction that determines economic dynamics and whatever complex features it may exhibit in time and space. The corresponding strategy simplifies the description of the small component unit, and simple agents/markets/institutions are made to interact to produce complex dynamic outcome. These two views have always been simultaneously present in economics. Their difference becomes relevant as soon as we leave the safe nest of the equilibrium-only analysis and extend our walk to the frontiers of out-of-equilibrium dynamics.

The dispute between the two views was never settled in favour of one or the other and this should be no surprise. (The same situation exists in the computing industry too, with the conflict between supercomputers and the philosophy of the web.) So far, neither modelling philosophy, too often taken to reflect pre-theoretical or ideological views, has proved superior. This book puts together examples: the macroeconomic chapters in Part II reflect mostly the former viewpoint, chapters in Part III the latter.

The conception of the economies as decentralised parallel structures of decision-making units has been traditionally considered the gateway to the dynamics of instability and fluctuations, in general what is now called complex dynamics. The classical theories of the business cycle as a disequilibrium phenomenon are children of this very idea, crystallised in R. Frisch's notions of structure and macrodynamics. An anticipatory analysis can be found in the contribution of Richard M. Goodwin on the dynamical coupling between sectors, where the mere phase interlocking between otherwise regular parallel oscillators is shown to be able to generate many sorts of complicated sectoral and aggregate time series. It may be useful to compare it with modern exercises on endogenous fluctuations and chaos where, instead, complex dynamics is basically the result of a non-linear model of a single-sector macroeconomy (thus, in the sense above, they show a *serial view*). But there is now also a vast literature on dynamics from interacting heterogeneous components, e.g. fundamentally economic agents, but also markets and sectoral growth cycles. The contributions appearing in this volume can be seen as a natural extension of the classical approach. Such models also try to cope with the issue of co-ordinating the disaggregated framework with aggregate outcomes, re-uniting micro and multisectoral with macroeconomic analyses. Does interaction *per se* generate growth, as well as, or as much as, fluctuations? Some of the contributions presented hereafter take up this point.

## Parts I and II

The interaction between economic fluctuation, growth and structural change is the main thrust in Part I which is dedicated to a survey of facts and interpretations. The strength of such an interaction is the fundamental motivation for Solomou's rejection of the 'methodology of stylised facts' searching for repetitive patterns and 'dynamical laws' of one sort or another. In an essay rich in both historical reconstruction and methodological reflections, the author highlights the complications involved: the varying patterns of causality/interdependence relations between the key factors of shorter term fluctuations, and the interlocking of oscillations at various frequencies with structural change. He points out that, contrary to common beliefs, the post-war experience of growth in the developed countries is fundamentally a unique string in a long time series of events. Thus, the author promotes a historical perspective as an alternative the methodology of stylised facts which yields anyway a rather unstable and thus interpretation-unreliable set.

Yoshikawa focuses upon the exemplary intermingling of growth and structural change in the post-war history of Japan. He criticizes in particular various existing interpretations of the rapid growth era, to propose a two-phase model whereby an early (1950–60) modernisation process of a Lewis-type dualistic economy is followed by a dramatic break and re-orientation. The change in household structure and urbanisation, driving labour force away from traditional agricultural employment and creating new urban demand, generated the unprecedented expansion of the earlier phase. This yielded, at the end of the era, to market saturation creating the need for a new, externally driven expansion. Domestic demand was substituted by foreign demand as an engine of growth. The chapter contains also interesting remarks on the prospects of NICS of various generations and on the experience of countries like Italy sharing with Japan a dualistic structure.

In the chapter by Böhm and Punzo a new framework is introduced where growth is naturally associated with fluctuations and structural change is defined as a qualitative change in the growth model, thus spanning a dynamic menu that is richer than expected. The chapter reviews in this light empirical findings for a set of European countries, the USA and Japan. It is shown that these went through repeated structural changes, but also that growth models were different across countries and were strung differently in each country's history. Observed high irregularity and cross-country variability reflect shock responses, of course, but more deeply they reveal the workings of the countries' own structures. Looking for changes in the growth model is one way of capturing structural change, and it seems to be the most natural one in the formalised setting in the dynamical systems style: it is a regime shift. The notion of regime dynamics translates naturally into that of Day's multi-phase dynamics, whose application to very long run growth is well illustrated in the chapter jointly written with Pavlov (opening Part II). Their Generalised Evolutionary Model (GEM) focuses upon the evolution of economic and social macrostructures through phases of growth following sometimes an irregular, by no means determined sequence, and it accounts for such an evolution through the working of an internal instability mechanim. The GEM is calibrated to reproduce (in the simulation sense) what is known of the history of mankind, an exercise which illustrates an approach to qualitative analysis aptly termed *qualitative econometrics*.

In fact, the output of a dynamic model exhibits scenarios, artificial histories in other words. In general, what matters in these simulation exercises cannot really be the quantitative coherence or closeness of the artificial to the actual time series, so that debate goes on about their uses. One proposal comes from the real business cycle approach, illustrated in the chapter by Prescott, where some quantitative criterion is retained to assess the model at hand against evidence. As an alternative, following Day and Pavlov, artificial histories can be treated as qualitative descriptions. Their worth is in their capability of reproducing the shifting across phases (or regimes) following a given known pattern. Thus, the model may explain to a certain extent features of actual macroeconomic behaviour; adding some extrapolation exercises, we get to qualitative econometrics.

Clearly, an economy whose history exhibits repeated regime or phase switches can hardly be described by an equilibrium technique. The analysis of scenarios of out-of-equilibrium dynamics is the centre of the neo-Austrian approach proposed by Amendola and Gaffard, highlighting the conditions for the emergence of co-ordination problems in the decision processes as well as within the production processes. Thus, stability towards a given state or attractor (that is the core of the conventional dynamical modelling) is no longer the key issue and, actually, it turns out to be not even definable as a property, given the continually shifting nature of the system's theoretical equilibrium. Instead, it is the issue of the sustainability of the dynamic process itself, i.e. the viability of the growth path, that comes to the fore redefining the very contours of economic policy.

The two chapters concluding Part II review and contribute novel insights into the out-of-equilibrium approach. Flaschel's contribution offers a disequilibrium framework where a monetary economy works according to rules and laws set out by a composite model built upon the seminal contributions of Keynes, Marx and Goodwin. This economy's dynamic outcome shows a strong bias towards fluctuation rather than steady growth, Flaschel building upon the alternative tradition that looks at the market economy through the pessimistic eyes of the classical thinker.

Iwai tackles the time-honoured issue of demonstrating the possibility of reconciling the notion of long run with the notion of disequilibrium. The test ground for his effort is a classical issue, the possibility of a long run state where profits are above their normal levels. The effort proves successful with the redefinition in statistical terms of the very notion of long run equilibrium, which now allows for a cross-firm distribution of technologies, as it results from the continuously on-going process of innovation. Baumol's contribution, on the other hand, looks at the long run growth aspects of that same process and offers an in-depth analysis of some of its microeconomic features. He explains why the evidence of unprecedented growth record of the market economy contradicts the poor performance predicted by traditional growth theory. The key argument is centred on the trade-off between flows of inventions and distribution of benefits, and therefore on the novel observation that positive spillovers (linked with non-appropriability) do not necessarily impede the innovation process. As, on the contrary, a certain amount of spillovers does fuel performance, the author gives a positive answer to the question previously raised, whether interaction can generate growth *and* at the same time fluctuations. This vindicates Schumpeter's (and Goodwin's) viewpoint.

## Parts III and IV

Three contributions in Part III, by Masanao Aoki, Hogg *et al.*, and Delli Gatti and Gallegati, share the common theme that dynamics is produced by interaction between heterogeneous agents. The first of them shows how the property of an interacting agents economy of exhibiting multiple equilibria states (with different stability properties) is essential to generate asymmetric fluctuations, but it also permits the analysis of equilibrium selection 'in the limit', i.e. when the size of the economy, defined by the number of agents, tends to infinite. Variability and fluctuations are obtained by Hogg *et al.* in an analytical context that recalls a classical theme in the theory of interacting markets. The obvious reference is to the studies on stability of general equilibrium, but also to the already mentioned paper by R.M. Goodwin on dynamical coupling. It is shown in an original manner that stability depends upon the degree of interdependence among markets, and the novel result lies in that such stability decreases with increasing dynamic interaction that results from economic agents' learning and broadening the environment in which they operate. Therefore greater integration, even in presence of locally weak connections, can only increase the likelihood of the fragility of the whole system. Recent turmoil on some key financial markets gives the appropriate background to appreciate the relevance of the conclusion or its 'realism'.

The dynamics driven by the actions of heterogeneous agents provides the topic of the contribution by Delli Gatti and Gallegati, which reviews a burgeoning literature on the structure of financial markets and its link with fluctuations. Here, the focus is on the heterogeneity among agents, a clear departure from the shortcut provided by the representative agent hypothesis, as well as on its effects on the outcomes predicted by a variety of macroeconomic models. Of course, given the location in the present book (and the inclinations of the authors), this review exhibits to a line of research where, instead of stabilising the economy towards a unique equilibrium, agents' heterogeneity generates complicated dynamics. Fluctuations can result from the distribution of the agents' characteristics acting as a powerful oscillatory propagation mechanism in the presence of even the smallest shock. One more door is thus opened to an endogenous explanation of observed fluctuations (the literature on endogenous oscillations, sunspots, complex dynamics and chaos having opened the first one). Putting all these contributions together, one seems to 'sense' that a new generation of models is now around where growth, fluctuations *and* structural change are simultaneously explained by the same theoretical engine, the structure of the economy.

Part IV moves onto more methodological issues. The opposition made by Prescott between the inductive approach that would be appropriate only to the natural sciences with the deductive one of economics (and perhaps other social sciences) meets the dual position of Rissanen questioning the foundations and practice of statistical modelling. The two authors, however, address different issues. Rissanen's position, which rests upon the notion of useful (and learnable) information, was not too long ago reflected by Sims, whose interpretation of the work of science in general as a search for laws as compressed data strings, shares the philosophy of Rissanen's Minimum Description Principle and in general of information theories in the grand tradition of Shannon and Kolmogorov. With a highly technical argument, which is however made sufficiently palatable to students versed in modern statistical techniques, Rissanen highlights the relevance of the notions of stochastic complexity for the analysis of given data strings and the recovery of candidate generating models. (Data strings are often time series to the economist, typically so in studies of dynamics.) Hence comes the link with Prescott's position where simulation output or predictions of the theory are to be compared (in a special way) with actual data. Prescott reviews this RBC methodology and indicates key issues still seeking satisfactory treatment, so that his chapter ends with a useful and stimulating research agenda.

Brock's chapter critically reviews the literature at the interface between theories of complex economic dynamics, which are typically non-stochastic, and the statistics of time series, attempting an assessment of gains from their cross-fertilisation, a project he himself has contributed to significantly. The impression is that the record is rather mixed, and that research is still going on (and, on big themes like this, it takes a long time to obtain new far-reaching results). This also indicates the nature and the scope of the theoretical research on complexity in economic dynamics. For its non-stochastic approach and its insistence on endogenous explanations, the latter has been taken to represent the alternative to the dominating paradigm, whereby dynamics is explained by fundamentals and off-equilibrium irregular paths result from stochastic disturbances. On the whole, the sober message in Brock's contribution is that a less partisan attitude is still the best the current researcher can entertain. One can connect this message with the main point in Hahn's contribution, that a lot of what was assumed to be endogenous was in fact exogenous, but a lot of the exogenous was assumed to be endogenous, too.

Sharp methodological demarcations as these may be useful to label theories and schools, but all theories should be understood first of all as experiments of a mental kind. Hence, the endogenous/deterministic explanation of fluctuations should be taken to show only how far one could go without calling in external forces and resorting to shocks. The same can be said of the distinction between exogenous and endogenous factors. Their survival, and recent revival, show the theoretical strength and appeal of such demarcations. This also reminds us that the frontiers of our economic knowledge have not yet been reached.

Lionello F. Punzo

## Acknowledgement

I wish to thank for their comments (with the usual caveats) J.-L. Gaffard and M. Puchet Anyul. Research for this preface and work in the final stages in the production of this book were carried on as part of a project on 'Recent developments in qualitative analysis and object-oriented simulation in economic dynamics', financed by the Italian MURST as Progetto di ricerca scientifica di rilevante interesse nazionale, 2000.

## References

Goodwin R.M. (1982), *Essays in Economic Dynamics*, London: Macmillan

Goodwin R.M. (1983), *Linear Economic Structures*, London: Macmillan

Goodwin, R.M. (1990), *Chaotic Economic Dynamics*, Oxford: Oxford University Press

Goodwin R.M. and L.F. Punzo (1987), *The Dynamics of a Capitalist Economy*, Cambridge: Polity Press, and Boulding, CO: Westview

Coricelli F., M. Di Matteo and F.H. Hahn (eds) (1998), *New theories in growth and development,* Basingstoke; New York: Macmillan

# Part I

# Facts and interpretations of growth and fluctuations

# 1 Economic cycles since 1870

*Solomos Solomou*

## 1.1 Introduction

Current research on business cycles has focused on the post-war period. The aim has been to derive the main 'stylised facts' of fluctuations. The approach here is very different. The emphasis is on the need for an historical perspective to business cycles. History inevitably provides us with a broader empirical basis that will allow us to formulate more general theoretical and empirical questions. It is argued that to focus on explaining the stylised facts of any one historical period to gain a general insight on economic fluctuations will lead to serious errors of interpretation.

An historical perspective is important not because history repeats itself but because history illustrates the evolutionary nature of business cycle behaviour and gives us an understanding of the factors that generate change. In economic systems that entail behavioural, institutional, structural and policy changes we can safely predict that the nature of business cycles will not be stable over time. The case for an historical perspective is founded on two important empirical features. First, the process of economic growth of modern economies has undergone massive structural change in the past two hundred years. One important aspect of structural change is the reallocation of production across sectors. For example, with industrialisation we would expect a change in the role of agricultural cycles in macroeconomic fluctuations. Given that these type of changes take place over a very low frequency, we need to observe business cycles over the long-term to gain insights on the resulting cyclical effects of structural change.

Second, policy regime changes have an observable effect on cyclical fluctuations. Business cycles during the rules-driven policy framework of the pre-1913 gold standard epoch were very different to those of the inter-war period. Similarly, the fluctuations of the Bretton Woods era were very different to those observed in the post-1973 era. Major policy regime changes have been few in number, once again, implying the need for a long-run perspective.

These observations raise serious doubts about the empirical relevance of the so-called 'stylised facts' approach to business cycles. This approach assumes that regularities exist over time and across countries (Lucas, 1981). However, many of these empirical regularities are usually derived from studies of very short time periods (often the post-1960 era) and a small selection of countries (usually the UK and the USA).

Even if we focus on a very limited set of empirical features we observe important changes. The average cyclical duration has changed over time. During 1870–1913 a number of variables (including aggregate investment, agricultural production and construction sector output) fluctuated with a long swing duration averaging about 20 years. During the inter-war period shorter fluctuations were observed. During the post-war 'golden age' the average cyclical period fell to 5 years. During the post-1973 era the average duration has once again lengthened, averaging approximately 10 years since the late 1970s. Cyclical amplitudes have also varied significantly over time. Low macroeconomic volatility during 1870–1913 gave way to high amplitude fluctuations during 1919–38; the stability of the post-war 'golden age' has been followed by the relatively more volatile post-1973 era. Patterns of co-movement of key variables have not been stable over time. Much of post-war research on business cycles has noted that prices and output have fluctuated contra-cyclically and has proceeded to explain this feature in terms of 'real' business cycle theory (Danthine and Donaldson, 1993). However, during the classical gold standard the relationship was not stable and during the inter-war period price and output fluctuations moved in a pro-cyclical manner (Cooley and Ohanian, 1991). *Assuming* universal stylised facts is not a realistic way of understanding business cycles.

This chapter surveys business cycle features across three historical epochs to illustrate some of the themes discussed in this introduction. Section 1 considers the pre-1913 epoch. Section 2 considers the changes observed in the inter-war period. Section 3 considers the post-war epoch.

## 1.2 1870 – 1913

The main institutional aspect of the period 1870–1913 is the gold standard. Between 1879–1913 Britain, France, Germany and America pegged their currencies to gold at a fixed rate. The relationship between this institutional-policy framework and economic fluctuations needs careful consideration. In particular, it is important to evaluate whether the rules-based policy framework modulated aspects such as the amplitude of fluctuations. The main structural aspect of the pre-1913 period is the large size of the primary sectors (such as agriculture and mining), as a percentage of the labour force and gross domestic product. Such a structure implies that supply-side shocks (such as weather shocks) are likely to have significant effects on sectoral and macroeconomic fluctuations. In terms of the world economy, the period is one of integration in trade, capital and labour flows between the industrial countries of Europe and the primary producing economies of the world. Such international linkages were important to determining the adjustment path to shocks in the industrial economies.

### 1.2.1 Describing cycles

Three different cycles have been identified by economic historians of the period: the Juglar trade cycle, the Kuznets swing and the Kondratieff wave. The Juglar cycle has an average period of seven to nine years. A number of studies have argued that this is the dominant cycle over this period (Aldcroft and Fearon, 1972; Lewis, 1978; Crafts, et al., 1989). For example, industrial production in the UK fluctuated with peaks in 1873, 1882, 1889, 1899, 1907 and 1913. The average peak to peak cycle is eight years. However, a number of data and conceptual problems should be noted before we accept this perspective to pre-1913 cycles. We should note a number of serious limitations in the historical industrial production series. For example the series for Britain (produced by Arthur Lewis) has been constructed using indicator variables and methods of extrapolation and interpolation. For a large proportion of the industrial production index Lewis has actually imposed a cycle of nine years on a priori grounds (Lewis, 1978; Solomou, 1994). A Juglar cycle is imposed on iron and steel products, commercial building, clothing, printing and chemicals, which account for over 28 per cent of Lewis' total industrial production index and 35 per cent of the manufacturing and construction index. Thus, the Lewis industrial production index provides only limited independent information for the importance of a Juglar cycle during this period. If we only consider the path of those industries where the cycle is not imposed a priori, the evidence for a dominant Juglar cycle is very weak. The construction sector sees long-term fluctuations of twenty years (Thomas, 1973). Coal production showed variations in trend without any discernible regular short cycle (Catao and Solomou, 1993). Investment was dominated by a long cycle of 23 years (Cairncross, 1953). Agricultural production was dominated by a long cycle of 20 years (Solomou, 1994). This evidence gives some support to Hicks (1982), who noted that during 1875–1914 business cycles became far more irregular in duration.

The Kuznets swing refers to a variation in economic growth that is longer than the Juglar trade cycle. The swings observed are variations either in levels or in rates of growth. The actual length of the swings found varies with different studies, but something between 14 to 22 years is representative. This type of fluctuation is most clear in describing the path of capital formation in the leading economies. British and American investment fluctuated along a 20–year cycle in the level of investment whilst French and German investment fluctuated along a long cycle in the rate of growth (Solomou, 1987). Irregular Kuznets swings can also be observed in the level and growth of agricultural production, construction output, migration, the sectoral terms of trade and trade balances (Solomou, 1987; Lewis, 1978; Thomas, 1973; Cairncross, 1953; Rowthorn and Solomou, 1991).

The Kondratieff wave is a cycle of prices and output with an approximate period of fifty to sixty years. In some of the earlier literature the Kondratieff wave has been used as a framework for understanding epochs such as the 'Great Depression' of 1873–96 and the *Belle-epoch* inflation of 1899–1913 (Kondratieff, 1935; Schumpeter, 1939). The empirical evidence suggests that we need to tread with care in the evaluation of Kondratieff waves. Most macroeconomic variables (such as GDP, investment and industrial production) have not followed a long-wave growth pattern (Solomou, 1987). However, long cycle adjustments in prices have been noted, particularly over the period 1873–1913 (Lewis, 1978; Rostow and Kennedy, 1979).

Given the historical discussion of multiple cycles it is useful to describe the relevance of this perspective using modern time-series methods. Solomou (1998) considers the empirical relevance of multiple cycle models by employing the Kalman filter to describe cycles in GDP for Britain, France, Germany and America. As can be observed in Figure 1.1a–1.1d the cyclical path of all these economies is depicted as the sum of short and long cycles. Although we have only presented the decomposition for GDP, multiple cycles are relevant at different levels of aggregation. This perspective provides an alternative to Hicks' (1982) description of irregular fluctuations. One aspect of irregularity is that the epoch was influenced by cycles of different average periods. Accepting the idea of multiple cycles as reality suggests one way of capturing some of the observed irregularity.

### *1.2.2 Rules-driven policy framework and business cycles*

The amplitude of business cycles was significantly lower during the gold standard period relative to the inter-war era (Sheffrin, 1988). The literature has attributed this to the rules-driven policy framework of the gold standard (Crafts and Mills, 1992). Such an idea offers only a partial explanation of this phenomenon. To understand the relatively low macroeconomic volatility of the period we need to consider the cyclical adjustment mechanisms operating over this period. Three adjustment mechanisms were of central importance. First, international labour mobility was exceptionally high. Second, capital mobility from Britain, France and Germany to the newly industrialising and primary producing economies created liquidity in the international economy. Third, capital flows sustained aggregate demand (by stimulating exports) during episodes of downswing in the capital exporting countries.

These features resulted in stabilising interactions between the domestic economy and the international economy. For example, depressed conditions in the national economy gave rise to a cyclical propagation mechanism that resulted in an adjustment of the labour market via overseas migration. International migration played an important role in accounting for cycles over this period (Thomas, 1973). The fall in home investment was also correlated with a rise of overseas investment and the income transfer overseas effected demand for the exports of industrial countries. Thus, compensatory equilibrating mechanisms were operating in the pre-1913 era. A rules-driven policy regime *by itself* was not sufficient to generate these cyclical features. It is the combination of historically unique circumstances with the gold standard policy framework that provided stabilising adjustment paths. Bayoumi and Eichengreen (1996) provide empirical evidence to support this conclusion: they find that the relative stability of the gold standard period cannot be explained by an absence of destabilising shocks; instead they show that there existed rapid adjustment to disturbances.

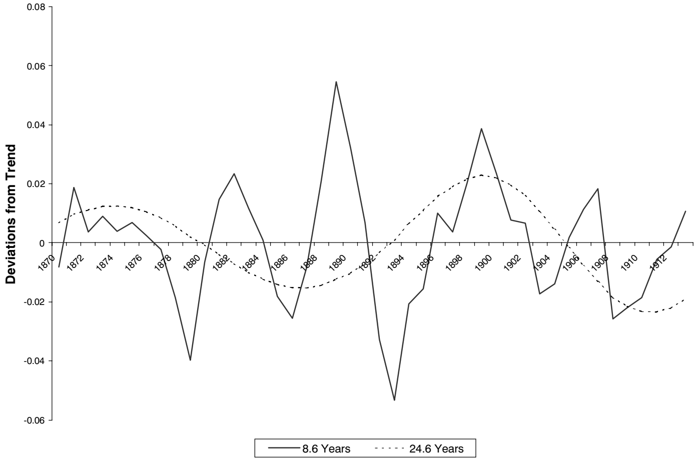


*Figure 1.1a* Kalman filter decomposition of UK GDP cycles, 1870–1913

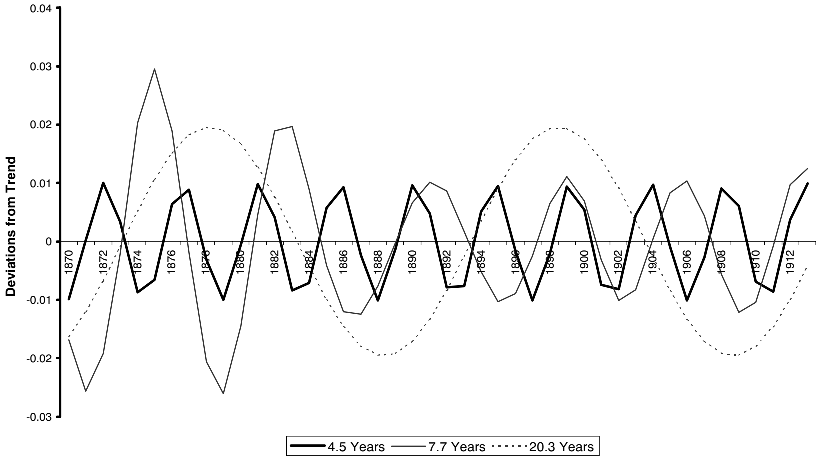


*Figure 1.1b* Kalman filter decomposition of French GDP cycles, 1870–1913

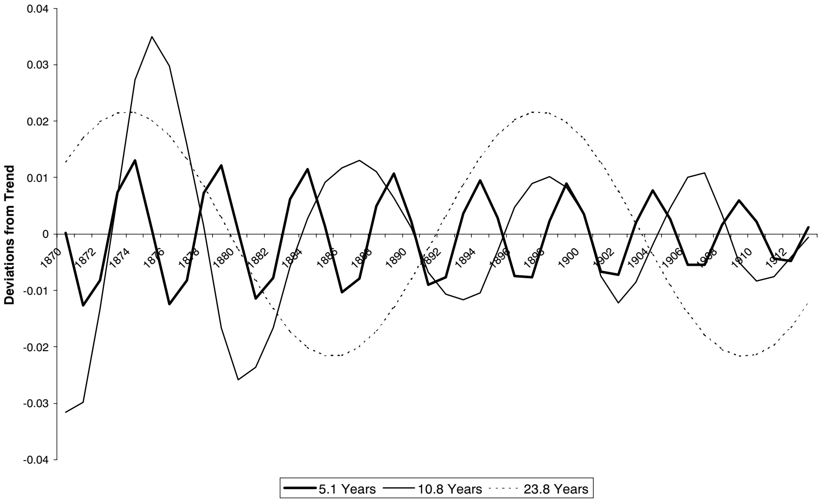


*Figure 1.1c* Kalman filter decomposition of German GDP cycles, 1870–1913

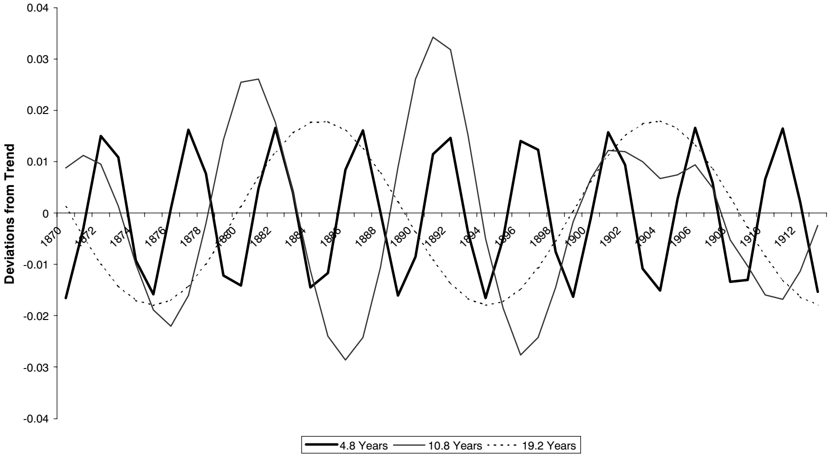


*Figure 1.1d* Kalman filter decomposition of US GDP cycles, 1870–1913

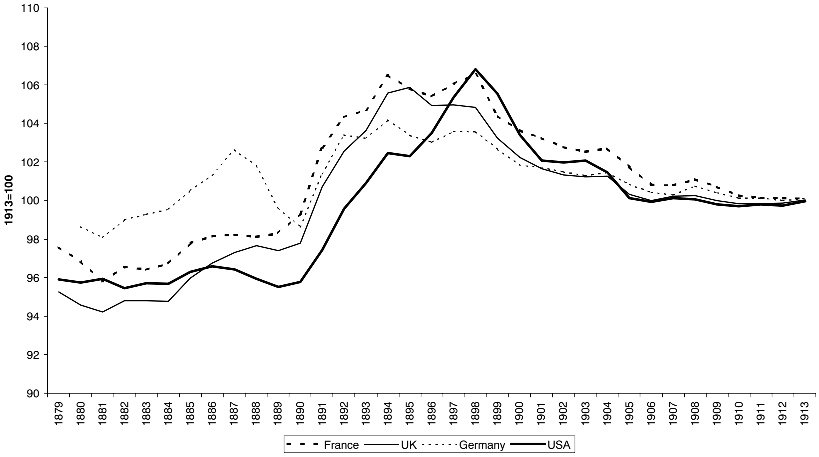


*Figure 1.2* Nominal effective exchange rates

Although the classical gold standard period is often thought of as a prime example of a rules-based policy framework, the non-universality of the system introduced significant nominal shocks. During the period 1879–1914, whilst all of the core industrial countries sustained a fixed exchange rate, the gold standard was not a global fixed exchange rate system. Alternative exchange rate regimes, such as the silver standard and paper currencies, prevailed in most primary producing countries – some of which witnessed considerable exchange rate variability (Ford, 1962; Nugent, 1973; Eichengreen, 1989; Bordo and Rockoff, 1996). Industrial and primary producing economies were not self-contained economic blocs. A large proportion of international trade during this period consisted of the exchange of industrial goods mostly produced in ‘core’ countries, for food and raw materials produced in ‘periphery’ economies outside the gold standard. Given these linkages, one would expect significant fluctuations in nominal effective exchange rates to have taken place in the ‘core’ countries.

Solomou and Catao (1998) calculate nominal and real effective exchange rates during this period to illustrate this effects. The nominal effective exchange rates of the four core economies are presented in Figure 1.2. The movements of nominal effective exchange rates in the core countries suggest the existence of stochastic trends, which would not be expected in a fully-fledged international gold standard. The magnitudes of the movements appear small when compared to the volatility of exchange rates during the inter-war period and the post-1973 era, but do stand out in a period when nominal variables were rather stable.[1] During the 1880s all four countries witnessed relatively low nominal effective exchange rates, relative to the whole period; this episode was followed by a phase of appreciation in the 1890s; the period from the late 1890s and early 1900s was one of depreciation. Thus, it is misleading to think of the era as one of rules giving rise to an absence of nominal shocks.

### *1.2.3 International economic cycles?*

One aspect of business cycles over this period that has been discussed in the recent literature is that the timing of fluctuations differed significantly across the major industrial countries (Backus and Kehoe, 1992; Eichengreen, 1994). At face value this suggests that national-specific shocks were more important than international business cycle transmission mechanisms. Table 1.1 reports the results of Backus and Kehoe (1992) which show that the only statistically significant correlation between different national GDP cycles is that between Germany and the USA, and that is small in magnitude.

The robustness and meaning of this result needs to be considered in the context of the methodologies used to derive it. The result is dependent on the kind of filter being used to decompose and describe economic cycles. Backus and Kehoe use the Hodrick–Prescott (H–P) filter which can be criticised for generating artefact cycles from the data (Harvey and Jaeger, 1993; Cogley and Nason, 1995). Moreover, the H–P filter focuses on deriving one short cycle from the data when, in fact, the pre-1913 data verify the existence of a number of cycles. The relevant question that needs to be addressed is, did countries share common cyclical paths at particular cyclical frequencies. The degree of integration of the international economy during the gold standard era was reflected in the existence of the type of international adjustments we have discussed above. Depressed investment and consumption opportunities in industrial countries were reflected in rising overseas investment, migration flows and a stimulus of the export sectors. This is an era when the long swing cyclical process is so pervasive across countries that it does not make sense to analyse business cycles as short fluctuations in GDP. The evidence reported above suggests that a multiplicity of cycles was the norm in the pre-1913 epoch. There is enough evidence to suggest that we can reject the idea that pre-1913 cycles were national-specific. The type of adjustments observed during the gold standard period resulted in significant interactions across countries. Most of these were observed over the low frequency fluctuations in migration, international capital flows and trade flows.

### *1.2.4 Weather shocks and business cycles*

The economic structure of pre-1913 economies suggests that significant fluctuation is arising from a coupling of the economic system with weather fluctuations as a source of supply-side shocks. The recent climatology literature has documented the existence of cyclical behaviour in weather variables such as temperature, rainfall (Lamb, 1977, 1982) and soil moisture deficits (Wigley and Atkinson, 1977). Such ‘cyclical’ and random weather variations affect the cyclical path of production in weather sensitive sectors, such as agriculture, construction and energy demand (Khatri *et al*., 1998; Solomou and Wu, 1999). What is the aggregate effect of such weather cycles? The accepted view of economic historians is that adverse weather ceased to be a significant shock to industrial economies. The formulation and dismissal of simplistic theories about weather and business cycles have also held

*Table 1.1* International output correlation: pre-1914

| | *UK* | *USA* | *Germany* | *Japan* |
|---|---|---|---|---|
| UK | – | | | |
| USA | 0.01<br>(0.14) | – | | |
| Germany | 0.03<br>(0.12) | –0.40<br>(0.13)* | – | |
| Japan | 0.08<br>(0.17) | 0.22<br>(0.15) | 0.14<br>(0.17) | – |

The entries show the contemporaneous correlation of cyclical variations of output. Numbers in parenthesis are standard errors: only the German–American correlation is significantly different from zero. * denotes significant at the 5 per cent level.

*Source*: Backus and Kehoe, 1992, p. 876.

back research in this area. For example, Jevons (1884, p.235) argued for a sunspot theory of the trade cycle. In general, modern economists dismiss such theories as unconvincing and misleading, although they continue to find favour in some of the literature (Zahorchak, 1983).

The impact of weather shocks will vary from sector to sector as the conditions favouring one activity may be adverse to another. The extent of such dampening will also depend on changes in the sectoral structure of the economy over time. The largest weather-sensitive sectors during this period were agriculture, construction and energy demand. In Britain, these three sectors accounted for approximately 25 per cent of GDP in the 1870s, falling to 17 per cent of GDP in the early twentieth century. In Germany and France these ratios were approximately 50 per cent of GDP in 1870, falling to 30–40 per cent by the early twentieth Century.

Solomou and Wu (2000) show that weather fluctuations had large effects on all these sectors. Using data from Britain, the effect range of weather shocks on the growth rate of construction and coal sectors was around ±0.05 during the period 1870–1913. A similar range is observed for the agricultural sector during 1880–1913, although this was significantly wider during the 1870s. The aggregate effect of these sector-specific effects depends on two features: first, the pattern of covariance of effects across different sectors and second, the relative weight of the different sectors in the macroeconomy and changes in these shares over time.

Consider the pattern of correlation of weather effects on different sectors (Figure 1.3 plots sector-specific weather effects). A significant positive correlation is observed between agriculture and construction sector effects (r=0.51)[2]. There is no evidence that weather shocks were having a neutral effect on the macroeconomy because adverse shocks to one sector were consistently being compensated by favourable effects to other sectors.

In order to evaluate the sum of all these sectoral effects the magnitude of the sectoral shocks are aggregated using GDP shares as weights. Figure 1.4 plots the shares of the three sectors in GDP. The agricultural sector is dominant in determining the weather effect on the economy. This arises because of the relative

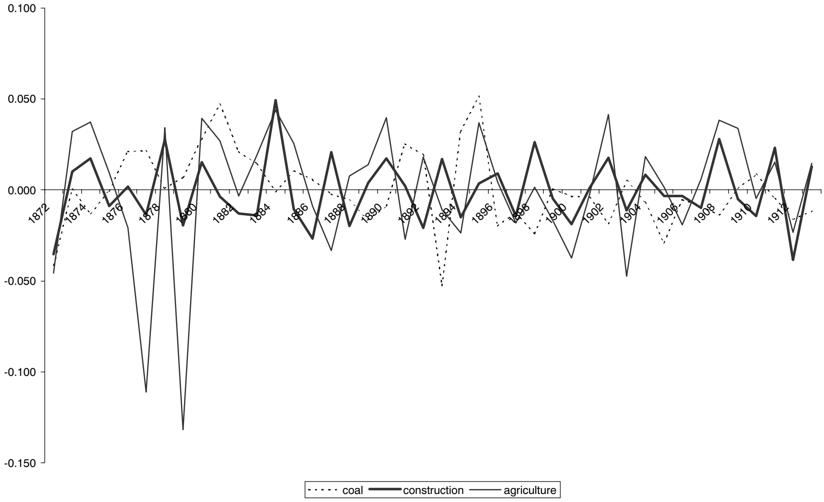


*Figure 1.3* Sectoral weather effects

size of the agricultural sector over this period, a pattern of positive co-movement between the weather effects affecting agriculture and construction and the wider effect range of weather shocks to agriculture. Figure 1.5 illustrates this result with a plot of the aggregate weighted weather effect and the weighted sectoral effects. It is clear that the total weather effect on the economy is dominated by the effect from agriculture and the co-movement with the weather effect on construction. The effect from coal simply adds some randomness to the magnitude of the total effect.

Aggregating sectoral weather effects using GDP shares as weights, suggests that the impact of weather shocks remained large throughout this period. The effect range of the sum of the sectoral shocks to GDP ranged from $+0.7$ per cent of GDP to $-1.5$ per cent of GDP. These estimates represent the lower bounds given the economic structure of other European countries compared to Britain. Moreover, because weather shocks are autocorrelated, we can safely conclude that weather added significant direct cyclicality to the economy. To dismiss the role of weather (even as late as the post-1870 period) in business cycles is ahistorical.

Summarising, economic cycles during the pre-1913 period are best depicted by multiple cycles. The causal structure behind such a process is likely to be complex. A number of adjustment processes were made possible by the nature of the world economy during this period. Free capital and labour movements across national boundaries resulted in long cycles in both these variables. The economic structure of the period also resulted in a coupling of weather cycles and sectoral fluctuations. At the macro level the relationship was not a simple coupling of one cycle with another. Instead, structural change acted as an important filter. Such structural change meant that agricultural fluctuations had a declining effect on the macroeconomy during the pre-1913 period, with the largest effects being observed

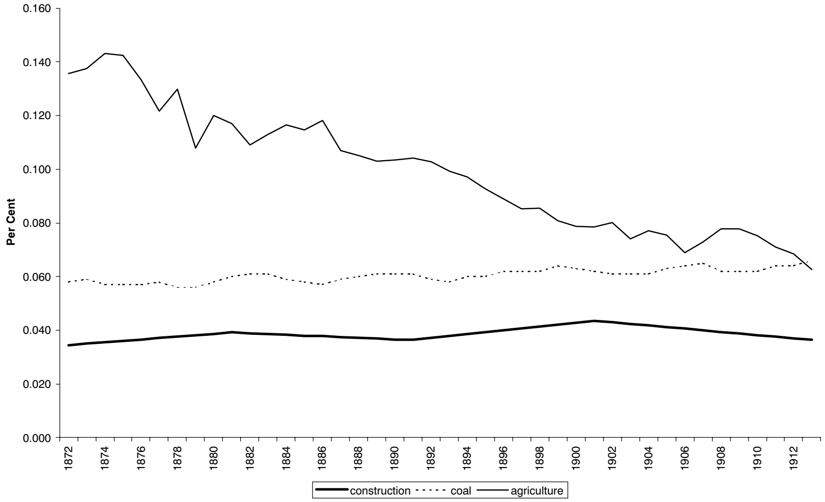


*Figure 1.4* Sectoral shares in GDP

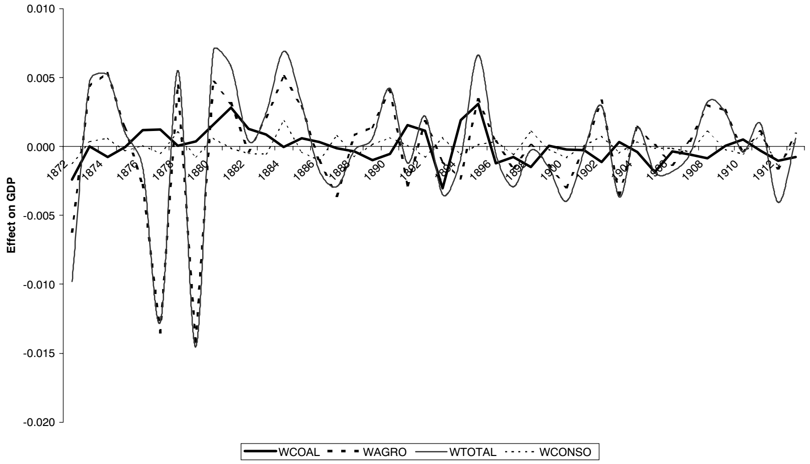


*Figure 1.5* Weighted sectoral effects and total weather effect

in the 1870s. An emphasis on the gold standard as the key feature determining fluctuations over the period is misleading.

## 1.3 Inter-war epoch

During 1919–38 the world economy witnessed economic fluctuations that were, in many ways, distinctly different from those of the past. A study of this period illustrates that whilst business cycles are a recurrent phenomenon their characteristics evolve and change over time. The following are some of the more important changes that emerged in the inter-war period:

### *Macroeconomic trend-stationarity*

During the pre-1913 gold standard era at least some of the major industrial countries displayed trend-stationarity in their *aggregate* growth paths, with a tendency to grow along a stable path in the long run (Mills, 1991; Solomou, 1998). This is the case for Britain and America over the period 1870–1913 and for Germany during 1880–1913. During the inter-war era these underlying growth paths were displaced in a persistently downward direction as a result of adverse business cycle shocks during the early 1920s and early 1930s. The timing of these adverse shocks varied by country: for example, Germany and Britain experienced persistent adverse effects in the early 1920s while America saw persistence arising from the adverse shocks of the early 1930s.

### *The 'passing of the kuznets cycle'*

The pre-1913 era (particularly during the period 1860–1913) witnessed Kuznets swings in the trends of a number of important variables, including GDP, migration, agricultural production, construction output, domestic and overseas investment, consumption, exports and the balance of trade (Solomou, 1987, 1998). During the inter-war period shorter cycles in all these variables became the norm, suggesting a break in the long swing growth process (Abramovitz, 1968).

### *Business cycle amplitudes*

Business cycle volatility was significantly higher in the inter-war than in the pre-1913 era (Sheffrin, 1988: Backus and Kehoe, 1992). The high variance of GDP and industrial production during the inter-war period is historically unique.

### *International cycles*

Business cycles showed high levels of international co-movement in the inter-war era relative to the pre-1913 era (Backus and Kehoe, 1992; Eichengreen, 1994). This high level of international co-movement stands out even when comparisons are made with the post-war era.

These new descriptive features raise a number of important questions: first, is the persistence of shocks influenced by the collapse of the rules-driven policy regime of the pre-1913 classical Gold standard, and its replacement by more discretionary policy regimes in the 1920s and 1930s? Second, why did the international adjustments of the long swing process (such as the inverse home and overseas investment swings) come to an end during the inter-war period, and what implications did this have for economic cycles? Finally, how do we account for the internationalisation of the business cycle during the inter-war period? These questions are addressed below.

### *1.3.1 Persistence of shocks*

Many business cycle shocks during the inter-war period had persistent effects on aggregate macroeconomic performance. This feature is in marked contrast to the pre-1913 gold standard era, which saw the workings of a number of long-run adjustment processes resulting in the feature of macroeconomic trend-stationarity. Examples of shocks that resulted in persistent effects are the 1920–1 depression in Britain, the German monetary stabilisation of 1923–4 and the Great Depression of 1929–33, which had long-term effects on a large number of economies.

Two key influences determined this new feature. First, the more discretionary policy framework of the inter-war period that arose out of the flexible exchange rate era (1919–25), combined with the attempts to re-establish the gold standard (in the form of the gold exchange standard) forced economies to deviate from their long-run paths on a permanent basis. This is very much at the heart of understanding the British and German experiences of the early 1920s (Solomou, 1998). The maladjusted gold exchange standard has also been given a central role in explanations of the 1929–33 depression (Eichengreen, 1992; Temin, 1989). The defence of the fixed exchange rate regime by most countries during 1929–31 enforced excessive monetary and fiscal deflation. In the USA this had a devastating effect on the stability of the financial system with one third of American banks failing between 1929–33. The collapse of the financial system, combined with a high debt overhang from the 1920s prevented a full revival of investment and consumer durable demand in the 1930s. Bernanke and James (1991) show that this mechanism is applicable to a wide number of countries. Countries that defended the gold standard for long periods raised the probability of financial crisis. Once financial crisis occurs, the event is likely to have persistence effects by raising the cost of credit and limiting its availability (Bernanke, 1983; Calomiris, 1993).

Second, the inter-war epoch did not have an equivalent set of international adjustment mechanisms to the pre-1913 era. The pre-1913 gold standard survived for so long partly because there existed viable international adjustment mechanisms to national-specific shocks: including international migration flows, trade protection and overseas investment (which stimulated the tradable sector). These adjustments manifested themselves in long swing fluctuations in a number of international economic variables, such as exports, overseas investment and international migration. The inter-war era saw an abrupt end of many of these adjustment outlets. Legislation in the New World prevented mass emigration as a solution to mass unemployment. The disintegration of world trade, partly due to protection policies and a collapse of overseas investment during 1928–38, prevented export growth from stabilising the effect of domestic demand shocks. Instead, severe business cycle shocks left economies with high unemployment and low output levels. The 'passing of the Kuznets cycle' (Abramovitz, 1968) in the inter-war era is of central importance to business cycle experiences, just as the presence of Kuznets swings before 1914 represented the workings of various stabilising cyclical adjustment mechanisms in the international economy.

### *1.3.2 Business cycle volatility*

The evidence of high output volatility raises a number of questions: did the severity of shocks increase in the inter-war relative to the past; is wage and price flexibility lower (forcing greater quantity adjustments on the economic system); did the transition to American economic leadership of the world economy transmit volatility to the rest of the world?

The nature of shocks did change significantly. Monetary policy shocks were of far greater importance in the inter-war period than before 1913 (Capie and Mills, 1991, 1992; Eichengreen, 1992). The flexible exchange rate era of 1919–25 was associated with a number of severe monetary and exchange rate variations; the malfunctioning gold exchange standard imposed excessive monetary deflation on the world economy between 1929–33 (Temin, 1989; Eichengreen, 1992); the devalued exchange rates of the 1930s made possible the use of discretionary monetary policy to stimulate economic recovery (Eichengreen and Sachs, 1985; Solomou, 1996).

Maladjustment and inflexibility in the labour market was emphasised by contemporaries of the inter-war period. However, although wage rigidity was present in the inter-war era it was also a feature of the pre-1913 era (Lewis, 1978). In an examination of the wage behaviour of Britain, France, Germany, the USA and Sweden, Phelps-Brown and Hopkins (1950) found that the degree of wage rigidity was comparable in the two periods in all these economies. More recent evidence suggests that only the USA provides an example of increased wage rigidity in the inter-war period (an outcome of the New Deal labour market policies in the 1930s) relative to the pre-1913 era.

Increased volatility during the inter-war period is an international phenomenon. Kindleberger (1983) has attributed this to the leadership structure of the world economy after the World War I. Although the USA emerged from the World War I as the only major capital exporter, it was not willing to take the lead in stabilising the international economy. Kindleberger sees the contra-cyclical British overseas investment before 1913 (captured by an inverse long-run relationship between home and overseas investment levels) as reflecting stabilising international behaviour by Britain. In contrast, the USA followed a pro-cyclical pattern of overseas investment that destabilised the world economy. While Kindleberger views these features as a reflection of national leadership qualities, it seems more plausible that the existence of 'Frontier' economies before 1913 made the absorption of capital by the world economy more likely. Given comparable profitability conditions, it is most likely that British investors would have responded in similar ways to their American counterparts during the inter-war period. The outcomes of the pre-1914 era were mainly fortuitous rather than planned; as in the inter-war period, these were the market outcomes of individuals seeking high rates of return[3].

### *1.3.3 International business cycles*

Much of the modern empirical literature on business cycles suggests that whilst pre-1913 fluctuations were national-specific, an international business cycle clearly manifested itself during the inter-war epoch (Backus and Kehoe, 1992; Eichengreen, 1994). As noted above, this result is sensitive to the definition of business cycles that have been employed by these studies. Backus and Kehoe use the Hodrick-Prescott (H–P) filter to derive cyclical decompositions. The working definition of business cycles that the H–P filter implies is one that emphasises patterns of high frequency fluctuations. Longer cycles in the data are assigned to the trend component.

Focusing on the co-variation of high frequency cycles misses relevant business cycle information. Pre-1913 fluctuations were highly integrated, but the strongest linkages are observed in the low frequency fluctuations. What changed in the inter-war period was not the creation of an international business cycle, but a shift in the period of the cyclical co-variations. The passing of the Kuznets cycle (in terms of international swings in migration, capital and export growth) modulated a different type of international business cycle. As large groupings of countries pursued common policy aims, such as the re-establishment of the gold standard in the 1920s, the defence of the gold standard parities during the early 1930s and the exit from gold in 1931–3, patterns of business cycles became conditioned by common policy regimes. This resulted in significant co-variations in the high frequency cycles across countries.

The new features of persistence, high volatility and high covariance of short cycles across countries created an important role for economic policy management. Whilst the economic environment of the pre-1913 period contained a number of adjustment mechanisms that activated a mean reverting tendency in the aggregate path of the economy, during the inter-war period shocks resulted in more persistent effects. Discretionary policy took on a new role of adjustment to shocks in the economy. Studies of the 1930s show that countries that devalued and pursued active expansive monetary policies saw a far better path of economic recovery during the period than those countries pursuing conventional policies, such as the gold bloc (Eichengreen, 1992; Temin 1989; Solomou, 1996)

## 1.4 Post-war cycles

The post-war period has witnessed a number of institutional changes that were to have a major effect on economic fluctuations. The Bretton Woods arrangements linked exchange rates to the dollar with the 'adjustable peg' mechanism, re-establishing rules-driven policy frameworks and breaking from the policy responses of the inter-war epoch. This system collapsed in the early 1970s and attempts were made to re-establish exchange rate stability within newly emerging trading blocs during the 1980s and 1990s. Comparing the period of the post-war 'golden age' with the post-1973 era allows us to evaluate the effect of different policy regimes on business cycle behaviour. The post-war period is also interesting in the light of two important long-run structural changes. First, the rapid growth of the government sector in most of the major industrial countries has been noted as a stabilising influence on aggregate demand, reducing the volatility of business cycle fluctuations relative to the past (Tobin, 1980; Zarnovitz, 1992). Secondly, the rapid growth of intra-regional trade has resulted in significant changes to the structure of world trade, and thus in the international transmission of shocks. These features have resulted in new business cycle features.

### *1.4.1 Business cycle amplitudes*

Table 1.2 shows that there has been a marked reduction in the variance of macroeconomic output fluctuations in the post-war era relative to the inter-war era. The volatility of the post-war period is also lower than for the pre-1913 era, suggesting that post-war policies and long run structural changes (such as the growth in the size of the Government sector) may have generated stabilising effects on modern economies (Burns, 1960; Tobin, 1980; Zarnovitz, 1992).

Some recent work has, however, questioned the comparability of these long-run data series. For example, Romer (1986) has argued that the high variance of the pre-1913 US economy is largely a statistical artefact, resulting from Simon Kuznets's estimation methods for US gross national product. Romer has revised the existing macroeconomic series on the assumption that they have an artificially high variance for the pre-1913 period. Simon Kuznets estimated an annual series of GDP using regression analysis. The regression series was constructed by taking the period 1909–38, a period when fairly reliable estimates of GDP can be obtained using the income–expenditure approach. Kuznets regresses the percentage deviation from trend of GDP on the percentage deviation from trend for aggregate commodity production.[4] He then uses the estimated coefficient to form an estimate of GDP for the period 1869–1918. However, the regression for the period 1909–38 is heavily influenced by the depression of the 1930s. Thus the constructed series may be biased to generate large cyclical variations. The percentage deviations from trend of GDP and commodity output move much closer to 1:1 during the 1930s than for other periods. During 1909–28 the coefficient is only 0.6. Romer (1986) uses the coefficient for the period 1909–28 as the basis for constructing the GDP series and this yields a substantially less volatile series before 1929. In the light of the revised data, Romer argues that the depression of the 1930s is exceptional and nothing comparable can be found before 1929. In fact the variance of US business cycles for the period 1870–1929 looks comparable to that for the post-war era, questioning the idea that demand management has moderated cyclical fluctuations after World War II.

There are, however, a number of problems to consider. If we assume that the macroeconomic relationships of the period 1870–1909 are similar to those for the period 1909–1928 then we observe a significantly lower amplitude than in the original Kuznets series. However, it is unlikely that these quantitative relationships will remain stable over such long periods. Moreover, the existence of long swings in the American economy suggests that Romer's revisions are unlikely to be correct because the period 1909–1928 is mainly capturing an upswing phase in the American economy; the behaviour of the 1930s, may be more relevant to depicting depression phases such as 1890s. Romer's assumptions are likely to have introduced artificial stability in the output data for the pre-World War I period.

Accepting the need to revise the US national income data Balke and Gordon (1989) have constructed a GDP series using new information on distribution, construction and consumer prices. Their results suggest that the variance of the pre-1914 economy is 1.77 times greater than for the post-war. Although this is lower than the original Kuznets data the results suggest that the pre-1914 economy was far more volatile than the post-war era, reinforcing the conventional picture.

Table 1.2 also shows that the phenomenon of a low variance during the post-war period is observed in a large number of countries, with very different historical data construction methods. It would be difficult to argue that all these long run changes are a statistical artefact (Backus and Kehoe, 1992; Sheffrin, 1988). Thus, although measurement errors make it difficult to compare long run historical data, the evidence is consistent with the idea of a structural shift towards relatively low business cycle volatility after World War II.

### *1.4.2 Cyclical duration*

Cyclical duration averaged 3 to 5 year growth cycles during the golden age of the 1950s and 1960s (Zarnovitz, 1992). Although short inventory cycles have been observed for the pre-war period, such a short duration for the major economic cycle is historically unique. The average length of the cycle has increased significantly since the 1970s. Analysing data for 1960–86 using the statistical methodology of the maximum entropy spectrum to determine cyclical duration for the OECD economies Hillinger (1992) found that cycles have increased in duration, although average duration varies significantly across countries. A large number of countries were influenced by a medium-term cycle of 12 to 15 years in duration.[5] The evidence suggests that cyclical durations have not been stable even within the post war era, with a discernible shift in average duration since the late 1960s. Although most business cycle theories do not seek to explain a regular periodic cycle, the observation of shifting mean cyclical duration suggests that the adjustment processes to shocks and/or the nature of shocks undergo significant change over time.

### *1.4.3 International synchronization*

The post-war business cycle was not a world economic cycle; instead bilateral cyclical linkages can be seen across a variety of countries. Strong bilateral co-

*Table 1.2*: Standard deviation (per cent) measures of output volatility (standard errors in parenthesis)

| *Country* | *Prewar* | *Inter-war* | *Post-war* |
|---|---|---|---|
| Australia | 6.30 | 4.85 | 1.93 |
| | (0.72) | (0.75) | (0.19) |
| Canada | 4.47 | 9.80 | 2.22 |
| | (0.43) | (1.40) | (0.23) |
| Denmark | 3.02 | 3.41 | 1.88 |
| | (0.22) | (0.64) | (0.20) |
| Germany | 3.35 | 10.19 | 2.30 |
| | (0.32) | (1.61) | (0.28) |
| Italy | 2.52 | 3.59 | 2.05 |
| | (0.24) | (0.46) | (0.17) |
| Japan | 2.42 | 3.13 | 3.11 |
| | (0.24) | (0.44) | (0.32) |
| Norway | 1.85 | 3.49 | 1.76 |
| | (0.16) | (0.65) | (0.17) |
| Sweden | 2.43 | 3.74 | 1.45 |
| | (0.37) | (0.59) | (0.12) |
| United Kingdom | 2.12 | 3.47 | 1.62 |
| | (0.24) | (0.37) | (0.21) |
| United States | 4.28 | 9.33 | 2.26 |
| | (0.38) | (1.27) | (0.18) |

*Source*: Backus and Kehoe (1992)

movements were observed between the UK and the USA, Japan and the USA, Canada and the USA, Germany and the UK and Canada and the UK. Insignificant linkages are observed between Japan and the UK, Japan and Germany and Japan and Canada. Such strong bilateral cyclical linkages are in marked contrast to the inter-war era when the business cycle is far more of an international phenomenon (Backus and Kehoe, 1992).

Using less formal statistical methods Zarnovitz (1981) compared the business cycle timings of the US, Canada, the UK, Germany and Japan within the National Bureau reference cycle framework during the period 1948–80. He also finds strong co-movements between different sets of countries, such as the US and Canada and the UK and Germany, rather than an international business cycle that affects all the major countries simultaneously.

Using quarterly output data for a larger cross-section of countries for the period 1959–89 Danthine and Donaldson (1993) document both bilateral and international business cycle linkages. At the bilateral level there are some very strong cyclical co-variations. For example, Germany shows significant co-movements with Austria, France, Italy, Switzerland and the UK; the UK has strong bilateral linkages with France and Germany. Aggregating the EC countries into one bloc we can also compare how each individual EC member varies with the EC bloc. The co-movements are very much weaker, suggesting that it is strong bilateral ties that are important in the intra-European cyclical linkages, not bloc behaviour.

Comparing the individual EC members with the USA, their linkages have been far stronger than with the EC bloc. The data also allow us to compare inter-bloc cyclical influences: comparing the cyclical variations of the EC, the USA and Japan we see strong positive co-movements across these three major economic zones. At this level of aggregation there is some evidence of international business cycle linkages.

### *1.4.4 Price-output cyclical co-movements*

In the light of developments in the monetary theory of business cycles (Lucas, 1977) and the real business cycle theory (Kydland and Prescott, 1982) the cyclical relationship between prices and output is central to business cycle research. Lucas simply *assumed* that prices move pro-cyclically with output as a stylised fact. Analysis of the post-war data suggests the very opposite, with price and output fluctuations moving contra-cyclically (Danthine and Donaldson, 1993).

However, evidence from other historical eras does not suggest stability of this relationship in the long run. During the classical gold standard period price–output fluctuations do not show any consistent positive or negative pattern; the inter-war period is chiefly characterised by pro-cyclical movements; the post-war phase up to the 1970s has been dominated by contra-cyclical movements. The most recent period suggests further changes. Pro-cyclical movements have been observed in the USA and the European economies in the 1980s and 1990s; as an example, the depression of the early 1990s in the major industrial economies resulted in the lowest inflation rates since the 1960s.

Pro-cyclicality, contra-cyclicality and non-cyclicality are all possible ‘stylised facts’, depending on the historical period being observed. The aims of theoretical frameworks should be to determine the reasons for the observed shifts in this relationship across different historical periods. Theoretical discussions that have sought to explain a universal stylised fact are, thus, misleading.

The evidence raises a number of theoretical and empirical questions that need to be considered further. For example, pro-cyclical price–output fluctuations arise in the inter-war and in some phases of the post-1973 epoch; both of these phases are periods of monetary policy discretion and unsettled exchange rate experiences. In contrast during the classical gold standard period and the Bretton Woods era the leading economies sustained fixed exchange rates, limiting national policy discretion. Thus, a *prima-facie* case can be established that the shifts in price–output movements are, at least partly, determined by the monetary-exchange rate policy regime. This does not mean that the price–output relationship is not influenced by real business cycle influences, such as the effects of technology and other supply-side shocks. What it does mean is that in some periods these effects are of secondary importance to the impact of the policy regime.

### 1.4.5 Disaggregated business cycle volatility

Investment, consumption and aggregate output co-move over the business cycle. This, in itself, may seem a rather intuitive and uninteresting result. However, a further stylised fact has a central role in business cycle discussions: it is often argued that there is a regular variance structure to the disaggregated data, with investment volatility being greater than aggregate volatility. The phenomenon of *consumption smoothing* by individuals, as predicted by the permanent income and life cycle hypotheses, implies that consumption is expected to be the least volatile component of aggregate demand.

Empirical studies generally accept this description of the structure of the Disaggregated data (Hillinger, 1992; Lucas, 1977; Kydland and Prescott, 1982). However, a number of important ‘outliers’ to this generalisation need to be noted. The US Great Depression of the 1930s cannot be understood without an explanation of consumption volatility far in excess of investment fluctuations (Temin, 1976 and 1989; Calomiris, 1993). Similarly, to account for the depression of the early 1990s in both the USA and Europe we need to be able to explain the observed consumption shifts (Blanchard, 1993; Hall, 1993). These two events in history carry much weight in terms of their impact on the economics and politics of market economies. The depression of the 1930s is historically unique in terms of its severity and the depression of the early 1990s is the most severe downturn for fifty years. To focus exclusively on investment volatility as an explanation of business cycle fluctuations is incomplete in that there is a selection bias towards neglecting information from major depressions. Thus, although consumption smoothing is an important empirical feature that needs to be integrated into business cycle research, *episodic* consumption volatility also needs to be integrated into a broader picture. Explaining these consumption shifts, or at the very least recognising their economic implications, is essential if we are to understand some of the major business cycle fluctuations of the twentieth century. Moreover, the financial deregulation which has been taking place since the 1980s means that consumption behaviour is likely to play an increasingly important role in business cycle fluctuations in the future.

## 1.5 Implications for business cycle research

This survey suggests that an historical perspective is an essential ingredient to an analysis of economic fluctuations. The ‘stylised facts’ approach to business cycles runs the risk of generating misperceptions. For example, in defining business cycles as high frequency annual fluctuations we risk neglecting relevant information with regard to the workings of cyclical adjustment mechanisms in the past (and in the future). The historical evidence suggests that long swings were a central feature of business cycle fluctuations in the past. To neglect this information will result in incomplete analyses. The long swings of the pre-1913 epoch created adjustment mechanisms that help us understand some of the key features of macroeconomic fluctuations within the period. Similarly the ‘passing of the Kuznets cycle’ helps us understand some of the changed features of the inter-war period. Utilising the total information set on fluctuation in the economy will result in a more informed analysis of economic cycles than if we simply use a working definition of business cycles that focuses only on high frequency fluctuations. The evidence we have suggests that we need to think of (at least) a three level decomposition: stochastic trend, low frequency long cycles and high frequency short cycles. Although we may choose to focus on cycles of a particular frequency (for whatever reason) we should always bear in mind that when a number of cycles may co-exist, the total cyclical information will also be useful.

Irregularity in business cycles seems to be a normal historical feature. Whether we focus on period, amplitude or patterns of co-movements we do not observe stable long run regularities that can be depicted as 'stylised facts'. Explaining irregularity opens up a number of challenges in theoretical research on business cycles, within both the impulse and propagation perspectives to business cycles. The examples considered in this survey suggest that policy regimes, long-run sectoral structural change, changing international inter-relatedness have had important effects. Understanding this will help in understanding that the future will be different to the present.

## Notes

1 Given the stochastic trends of nominal effective exchange rates (EER) we investigated the existence of a common stochastic trend using the Johansen co-integration framework. The nominal EER movements of the core countries are found to be co-integrated over this period.
2 The correlations between construction and coal and agriculture and coal are $-0.14$ and 0.001 respectively
3 One aspect of inter-war international capital flows, which may have destabilised the world economy, is the link between capital flows and the 'recycling' of reparation payments (Solomou, 1998).
4 Commodity production is a composite index of industrial and agricultural production.
5 Most countries were also influenced by a short cycle 5 to 10 years and a shorter inventory cycle.

## References

Abramovitz, M. (1968) 'The Passing of the Kuznets Cycle', *Economica*, XXXV, 349–67.

Aldcroft, D.H. and Fearon, P. (eds) (1972) *British Economic Fluctuations*, 1790–1939, London.

Backus, D.K. and Kehoe, P.J. (1992) 'International Evidence on the Historical Properties of Business Cycles', *American Economic Review*, 82, (4), 864–88.

Balke, N.S. and Gordon, R.J. (1989) 'The Estimation of Prewar Gross National Product: Methodology and New Evidence', *Journal of Political Economy*, 97, 1, 38–92.

Bayoumi, T. and Eichengreen, B. (1996) 'The Gold Standard and the International Monetary System', in T. Bayoumi, B. Eichengreen and M.P. Taylor *Modern Perspectives on the Gold Standard*, Cambridge: Cambridge University Press.

Bernanke, B.S. (1983): 'Nonmonetary Effects of the Financial Crisis In the Propagation of the Great Depression', *American Economic Review*, 73, 257–76.

Bernake, B.S. and James, H. (1991) 'The Gold Standard, Deflation and Financial Crisis in the Great Depression: An International Comparison', In R.G. Hubbard (ed.), *Financial Markets and Financial Crises*, Chicago: University of Chicago Press.

Blanchard, O. (1993): 'Consumption and the Recession of 1990–1991', *American Economic Review*, 83, 2, 270–4.

Bordo, M. D. and H. Rockoff (1996) 'The Gold Standard as a Good Housekeeping Seal of Approval, *Journal of Economic History*, 56, (2), 389–426.

Burns, A.F. (1960) 'Progress Towards Economic Stability', *American Economic Review*, 50, March, 1–19.

Cairncross, A.K. (1953) *Home and Foreign Investment, 1870–1913: Studies In Capital Accumulation*, Cambridge: Cambridge University Press .

Capie, F. and Mills, T.C. (1991) 'Money and Business Cycles in the US and U.K., 1870 To 1913', *Manchester School*, 53, (Supplement), 38–56.

Capie, F. and Mills, T.C. (1992) 'Money and Business Cycles in the United States 1870–1913: A Re-examination of Friedman and Schwartz' *Explorations in Economic History* 29, 3, 251–73.

Cogley, T. and Nason, J.M. (1995) 'Effects of The Hodrick and Prescott Filter on Trend and Difference Stationary Time Series: Implications for Business Cycle Research', *Journal of Economic Dynamics and Control*, 19, 253–78.

Catao, L.A.V. and Solomou, S.N. (1993) 'Business Cycles During the Gold Standard, 1870–1913', *Department of Applied Economics Working Paper*, 9304, Cambridge University.

Calomiris, C.W. (1993) 'Financial Factors In the Great Depression', *Journal of Economic Perspectives*, 7, (2) 61–85.

Cooley, T.F. and Ohanian, L.E. (1991) 'The Cyclical Behaviour of Prices', *Journal of Monetary Economics*, 28, 25–60.

Crafts, N.F.R., Leybourne, S.J. and Mills, T.C. (1989) 'The Climacteric In Late Victorian Britain and France:A Reappraisal of Evidence', *Journal of Applied Econometrics* 4, (2), 103–18.

Crafts, N.F.R. and Mills T.C. (1992) 'Economic Fluctuations, 1851–1913: A Perspective Based On Growth Theory', In S.N. Broadberry and N.F.R. Crafts (eds) *Britain In The World Economy 1870–1939*, Cambridge.

Danthine, J.P. and J.B. Donaldson (1993) 'Methodological and Empirical Issues In Real Business Cycle Theory', *European Economic Review*, 37, (1), 1–36.

Eichengreen, B.J. (1989) 'The Political Economy of The Smoot Hawley Tariff', *Research In Economic History*, 11, 1–44.

Eichengreen, B.J. (1992) *Golden Fetters: the Gold Standard and the Great Depression, 1919–1939*, Oxford: Oxford University Press.

Eichengreen, B.J. (1994) 'History of the International Monetary System: Implications For Research In International Macroeconomics and Finance', In F. Van Der Ploeg, F. *The Handbook of International Macroeconomics*, Oxford.

Eichengreen, B.J. and Sachs, J. (1985) 'Exchange Rates and Economic Recovery In the 1930s', *Journal of Economic History*, 45, 925–46.

Ford, A.G. (1962) *The Gold Standard 1880–1914: Britain and Argentina*. Oxford: Clarendon Press.

Hall, R.E. (1993) 'Macro Theory and the Recession of 1990–1991', *American Economic Review*, 83, (2), 275–9.

Harvey, A.C. and Jaeger, A. (1993) 'Detrending, Stylized Facts and the Business Cycle, *Journal of Applied Econometrics*, 8, (3), 231–47.

Hicks, J.R. (1982) 'Are there Economic Cycles?', In J.R. Hicks, *Money, Interest and Wages: Collected Essays On Economic Theory*, Vol. II, Oxford.

Hillinger, C.(ed.) (1992) *Cyclical Growth In Market and Planned Economies*, Clarendon Press, Oxford.

Jevons, W.S. (1884) *Investigations in Currency and Finance*, London: Macmillan.

Kindleberger, C.P. (1983) *The World In Depression,1929–1939*, Harmondsworth: Penguin.

Khatri, Y.J., Solomou, S.N. and Wu, W. (1998) 'The Impact of Weather on UK Agricultural Production, 1867–1913',*Research in Economic History*, 18, 83–102.

Kondratieff, N.D. (1935) 'The Long Waves in Economic Life', *Review of Economic Statistics*, (1935), XVII, (6), 105–115.

Kydland, F.E. and Prescott, E.C. (1982) 'Time To Build and Aggregate Fluctuations', *Econometrica*, 50, (6), 1345–70.

Lamb, H.H. (1977) Climate: Present, Past and Future, London: Methuen.

Lamb, H.H. (1982) *Climate, History and the Modern World*, London: Methuen.

Lewis, W.A. (1978) *Growth and Fluctuations 1870–1913*, London: Allen and Unwin.

Lucas, R.E. (1977) 'Understanding Business Cycles', In K. Brunner and A.H. Meltzer (eds), *Stabilisation of the Domestic and International Economy*. Carnegie-Rochester Conference Series On Public Policy 5, Amsterdam and New York: North Holland.

Lucas, R.E. (1981) *Studies In Business Cycle Theory*, Oxford: Blackwell.

Mills, T. (1991) 'Are Fluctuations In UK Output Transitory Or Permanent?', *The Manchester School*, LIX, (1), 1–11.

Nugent, J.B. (1973) 'Exchange Rate Movements and Economic Development in the Late Nineteenth Century', *Journal of Political Economy*, 81, (5), 1110–35.

Phelps Brown E.H. and Hopkins, S.V. (1950) 'The Course of Wage-Rates in Five Countries, 1860–1939, *Oxford Economic Papers*, 2, (2), 226–96.

Romer, C. (1986) 'New Estimates of Prewar GNP and Unemployment', *Journal of Economic History*, 2.

Rostow, W.W. and Kennedy, M. (1979) 'A Simple Model of the Kondratieff Cycle', in Paul Uselding (ed.),
em Research in Economic History, 4.

Rowthorn, R.E. and Solomou, S.N. (1991) 'The Macroeconomic Effects of Overseas Investment On The UK Balance of Trade, 1870–1913', *Economic History Review*, XLIV, (4), 654–64.

Schumpeter, J.A. (1939) *Business Cycles: A Theoretical, Historical and Statistical Analysis of the Capitalist Process*, New York: McGraw Hill.

Sheffrin, S.M. (1988) 'Have Economic Fluctuations Been Dampened? A Look At Evidence Outside The United States', *Journal of Monetary Economics*, 21, 73–83.

Solomou, S.N. (1987) *Phases of Economic Growth, 1850–1973: Kondratieff Waves and Kuznets Swings*, Cambridge: Cambridge University Press .

Solomou, S.N. (1994) 'Economic Fluctuations, 1870–1913', In R. Floud, and D.N. McCloskey (eds), *The Economic History of Britain Since 1700*, Vol. 2, Second Edition, Cambridge.

Solomou, S.N. (1996) *Themes In Macroeconomic History: The UK Economy 1919–1939*, Cambridge University Press.

Solomou, S.N. (1998) *Economic Cycles*, Manchester University Press.

Solomou, S.N. and Catao, L.A.V. (1998) 'Effective Exchange Rates, 1879–1913', *Department of Appied Economics Working Paper 9814*, Cambridge University.

Solomou, S.N. and Wu W. (1999) ‘Weather Effects on European Agricultural Output, 1850–1913’, *European Review of Economic History*, December, 351–73.

Solomou, S.N. and Wu W. (2000) ‘Macroeconomic Effects of Weather Shocks 1870–1913’, *Department of Appied Economics Working Paper*, Cambridge University.

Temin, P. (1976) *Did Monetary Factors Cause The Great Depression?*, New York: W.W. Norton.

Temin, P. (1989) *Lessons From The Great Depression*, Cambridge, Mass: MIT Press.

Thomas, B. (1973) *Migration and Economic Growth*, Revised Edition, Cambridge: Cambridge University Press.

Tobin, J. (1980) *Asset Accumulation and Economic Activity: Reflections On Contemporary Macroeconomic Theory*, Chicago University Press, Chicago.

Wigley, T. M. L. and Atkinson, T. C. (1977) ‘Dry Years in South-East England since 1698’, *Nature*, 265, (5593), pp. 431–34

Zahorchak, M. (1983) *Climate: The Key to Understanding Business Cycles*, Linden, N.J.

Zarnowitz, V. (1981) ‘Business Cycles and Growth’, Reprinted As Chapter 7 of V. Zarnowitz (1992) *Business Cycles: Theory, History, Indicators and Forecasting*, University of Chicago Press, Chicago and London.

Zarnowitz, V. (1992) *Business Cycles: Theory, History, Indicators and Forecasting*, University of Chicago Press, Chicago and London.

# 2 Growth and fluctuations

## The post-war Japanese experience

*Hiroshi Yoshikawa*

## 2.1 Introduction

The century-long experience of the Japanese economy provides economists with a great opportunity to study growth, business cycles and structural change. Japan, so poorly endowed with natural resources, has grown to become one of the major industrial nations. Figure 2.1 shows annual growth rates of real GDP for the pre-war period from 1885 to 1940 and the post-war period from 1955 to 1995. The pre-war Japan kept 3 per cent growth on average for more than half a century, and the growth rate accelerated to almost 6 per cent in four decades after World War II. The standard Ramsey model cannot satisfactorily explain this postwar acceleration of growth.

Her growth has not been necessarily smooth. A glance at the figure reveals that the pre-war Japanese economy was on the whole more unstable than the post-war economy. The standard deviation of the growth rate is, in fact, 3.5 per cent for the whole post-war period whereas it is 3.7 per cent for the pre-war period (Table 2.1). The standard deviation for the post-war period is, however, inflated by the fall of growth rate in the early 1970s. If we divide the sample period taking into account this break, we find that the standard deviation declines to 2 per cent as against 3.7 per cent for the pre-war period. Why was the post-war Japanese economy more stable than the pre-war economy? It is an interesting question. But in this chapter, I simply draw attention to this important fact, and focus on the post-war period. A good general reference for the post-war Japanese economy is Nakamura (1981).

The Japanese economy started from ashes after the Second World War. Real GDP fell to a half of the pre- war peak level, and the economy suffered from hyperinflation. Most Japanese economists believe that the economy completed the

*Table 2.1* Growth rate of real GDP (%)

| | *1886-1940* | *1956-95* | *1956-70* | *1970-90* | *1990-95* |
|---|---|---|---|---|---|
| (1) Mean | 3.2 | 5.7 | 9.2 | 4.4 | 1.9 |
| (2) Standard Deviation | 3.7 | 3.5 | 2.1 | 2.3 | 2.0 |
| (3) (2)/(1) | 1.16 | 0.61 | 0.23 | 0.52 | 1.05 |

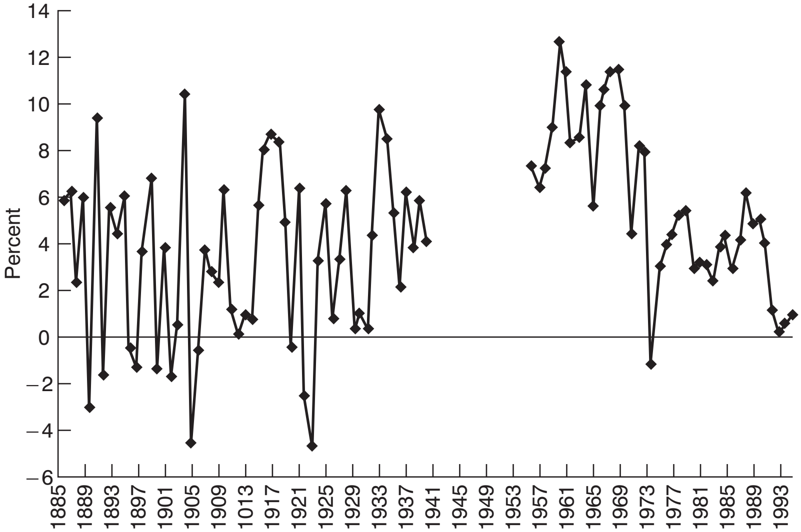


*Figure 2.1* Growth rate of real GDP

War Reconstruction by 1955, and, therefore, I take 1955 as a starting point of my study. Readers interested in the turbulent decade after the war are referred to Yoshikawa and Okazaki (1933) and references cited therein. It is convenient and standard to divide the post-war period into the following three sub-periods:

1. 1955–73: The average growth rate ten per cent
2. 1975–90: The average growth rate four per cent
3. 1990– : The average growth rate below two per cent

It is a central theme of this chapter to explain what mechanism generated extraordinarily high growth for such a long period as almost two decades after 1950. The era of high growth ended in the early 1970s. This ‘structural change’ of the Japanese economy is clearly seen in Figure 2.1. An obvious question is what caused this ‘structural change’. A standard explanation is the first oil shock in 1973–74 (Bruno and Sacks (1985), for example). In this chapter, I underplay the role of the first oil shock, and instead emphasize the importance of domestic factors. For the 1970s and 1980s, I emphasize the role played by exports and the exchange rate. Finally I briefly touch on the long recession in the 1990s.

## 2.2 The period of rapid economic growth and its end

The Japanese economy enjoyed an average of 10 per cent growth for almost two decades beginning in the mid-1950s. There are several factors which are believed to have contributed to the rapid economic growth of the Japanese economy during the 1950s and 1960s. The abundance of importable foreign technology is often mentioned as such a factor; Lincoln (1988) underlines the importance of this factor. However, it is not obvious that the stock of available foreign technology was much greater in the 1950s and 1960s than in the 1920s and 1930s or for that matter, in the nineteenth century. One might plausibly expect that such a stock was greatest when Japan opened her door to the West in the late nineteenth century.

A sharp decline in natural resource costs is also often mentioned. Technical progress in marine transportation is believed to have contributed to this effect. A sharp decline in marine transport costs has made natural resources commodities rather than part of the 'factor endowment' of individual countries in the post-war era. It naturally gave a leverage to the resource-poor Japanese economy. Total factor productivity in the industry, in fact, grew at the annual rate of 10 per cent in the post-war era as against 2–3 per cent in the pre-war period. Granted that these factors are very important I suggest below that (i) demographic trends, and (ii) diffusion of consumer durables, namely 'catch up of demand' played a particularly important role in the process of rapid economic growth.

### 2.2.1 The mechanism generating rapid economic growth

Recall that the Japanese economy in the 1950s and 1960s was a two-sector economy consisting of a rural agricultural sector and an urban manufacturing sector. As of 1950, nearly half of the total labour force was still engaged in agriculture (Table 2.2). In this respect, only Italy was comparable to Japan. Population continuously flowed from the former into the latter in the process of economic growth. The dual structure of the economy enabled the manufacturing sector to hire enough labour at the level of real wages which, determined in the agricultural sector with 'disguised' unemployment, were lower than the marginal product in the industrial

*Table 2.2* Total employment in primary, secondary, and tertiary occupation: Japan, USA, UK and West Germany, 1950 and 1989 (%)

| | *Primary (agricultural and mining)* | *Secondary (manufacturing and construction)* | *Tertiary (services)* |
|---|---|---|---|
| 1950 | | | |
| Japan | 48.3 | 21.9 | 29.7 |
| USA | 12.4 | 35.3 | 49.7 |
| UK | 5.1 | 47.5 | 47.0 |
| Germany | 23.2 | 42.2 | 32.4 |
| Italy | 46 | 27 | 27 |
| 1989 | | | |
| Japan | 7.6 | 33.8 | 58.1 |
| USA | 2.9 | 25.9 | 71.1 |
| UK | 2.4 | 29.0 | 68.6 |
| Germany | 4.9 | 38.4 | 55. |

*Note:* For Italy, 1951, UK and Germany, 1987.

sector. Growth of the manufacturing sector, therefore, entailed high profits in the same sector rather than an increase in real wages. The high profits in turn were supposed to induce high investment. All this is, of course, what Lewis (1954) describes as a typical process of growth of the underdeveloped dual economy.

The Lewisian model has been indeed successfully applied to the century long development of the Japanese economy by a number of economists (see, for example, Ohkawa and Rosovsky (1973), Minami (1968), and Inada, Sekiguchi and Shoda (1992)). In the Lewisian model, however, population flow between two sectors is taken solely as a *result* of the growth of the modern manufacturing sector. Minami (1968, 1970) and others demonstrate that internal migration was in fact quite sensitive to growth of the manufacturing sector; More people left rural agricultural areas for urban industrial cities in booms and *vice versa*. Yearly fluctuations of population flow was therefore a result of industrial growth. In the Lewisian model, the key factor behind this industrial growth is low real wages made possible by the existence of disguised unemployment in the agricultural sector.

In contrast to the Lewisian model, however, in what follows I argue that in the case of the *post-war* Japanese economic growth (1955–70), population flow between the two sectors was in fact the major factor in generating high demand for products of the industrial sector. In my view, population flow was a cause as well as a result of economic growth.

According to the Lewisian theory, population is supposed to continuously flow from the rural agricultural sector to the urban industrial sector, as actually happened in Japan. Among Asian developing countries, however, this Lewisian population flow occurred to a substantial extent only in a few NIES countries. The basic problem of the theory is that the growth of a modern industrial sector is sustained by high profits which are guaranteed by repressed real wages while demand for products in such a modern industrial sector is assumed to automatically emerge as production grows. In reality, demand does not emerge automatically. Demand is precisely the point which divides pre-war and post-war Japan.

The post-war Japanese growth during the period 1955–70 was led by domestic demand. For example, the contribution of net export to growth was on average –0.2 per cent for the high growth period. In the process of domestic demand-led growth, population flow and household formation played a crucial role. Because of the large-scale population flows (see Figure 2.2), *household formation* dramatically accelerated during the period of high economic growth, 1955–70 (Figure 2.3). I underline the fact that in this period, the growth rate of households forms a hump shape at a high level parallel to the growth of real GDP while the growth rate of population was quite stable at a much lower level of about 1 per cent . Population growth or the growth of the labour force which plays such an important role in the standard growth theory has little explanatory power for the high growth of the Japanese economy during the 1950s and 1960s.

As one might expect, traditional three generation merged households hardly increased during this period. Instead, the ‘core’ households consisting of a married couple, possibly with unmarried children, and an unmarried adult dramatically increased, particularly in urban industrial areas. When three generations of family members kept a traditional single household in rural villages, they would have

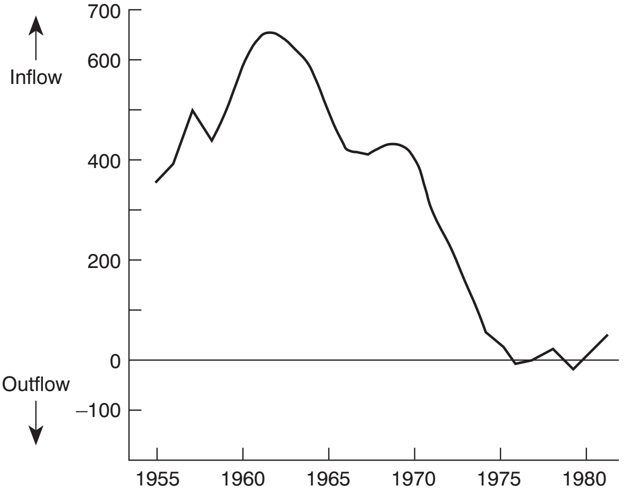


*Figure 2.2* Population flow 1955–80 into and out of Tokyo, Osaka and Nagoya metropolitan areas.

*Source: Annual Report on Internal Migration* Statistical Bureau, Management and Coordination Agency.

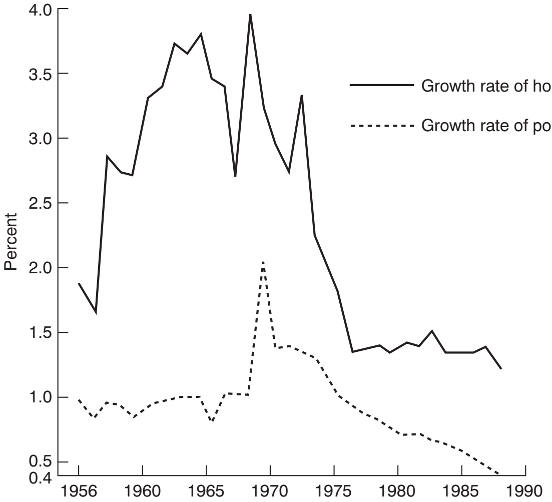


*Figure 2.3* Growth rates of households and population, 1956–90

Note: The figure for 1970 is an outlier; Okinawa prefecture (Ryuku Island) was returned to Japan by the United states in 1970.

needed only one of each consumer durable such as a refrigerator, television set, washing machine and car. In fact, young people giving up agriculture left rural villages for urban industrial areas where they formed new households. Household formation necessarily generated additional demand for houses, consumer durables, and electricity. In this way, population flow sustained high domestic demand in the period of high economic growth, 1955– 70.

Along with the creation of a large number of households, the high growth period was also the diffusion process of newly available consumer durables. The diffusion of consumer durables was facilitated by a steady decline in prices of those products over time on the one hand, and an increase in income on the other. Electric washing machines, for example, first appeared in the Japanese market in 1949. At the time, a machine cost 54 thousand yen while the average annual labour income was about 50 thousand yen. Understandably, only twenty machines were sold per month! By 1955, only six years later, however, the price of a washing machine had been reduced to 20 thousand yen while the average annual income had risen to above 200 thousand yen. About a third of households could afford to own a washing machine in 1955. The same story holds for other consumer durables as well. Since it was urban cities that led this diffusion process, urbanization not only created new households but also in itself sustained high demand for those consumer durables. By the end of the 1960s, however, most of the then available consumer durables saw saturation in the domestic market.

This whole process of domestic demand-led high economic growth (1955–70) is schematically summarized in Figure 2.4. Channels 1, 2 in the diagram have been well recognized: capital accumulation in the industrial sector raising labour demand brings about population flow from rural agricultural areas to urban cities. In addition to these well-recognized channels, I emphasize the oft-neglected and yet very important fact that such population flow in turn, creating new households and raising demand for consumer durables and electricity, ultimately sustained profitability of investment in manufacturing industry (channels 3, 4, 5). I underline that the role of newly available consumer durables was not confined to demand for those products themselves. Through an input-output interrelationship, they augmented demand for intermediate goods such as steel and electricity, and accordingly high investment in those sectors.

In this virtuous circle for high economic growth, low real wages were not so instrumental as Lewis (1954) emphasized. Rather it was growth of domestic demand that sustained profitability of investment. And for growth of consumption demand, a steady rise in real wages, rather than low repressed wages, is a contributing factor. In fact, in the pre-war period, except for 1920–21 and 1929–30, real wages saw little increase, while in the post-war period they enjoyed steady growth. A steady rise in real wages sustained effective demand for the post-war Japanese economy because the key product was consumer durables which not yet being international competitive, had to find a domestic market.

It is important to note that the pre-war Japanese economy was, like today's NIES, a typical export-led economy (see Shinohara (1961)). The key industry was cotton, and low real wages were in fact instrumental for international competitiveness. The export/GNP ratio in pre-war Japan was, for example, 20 per cent

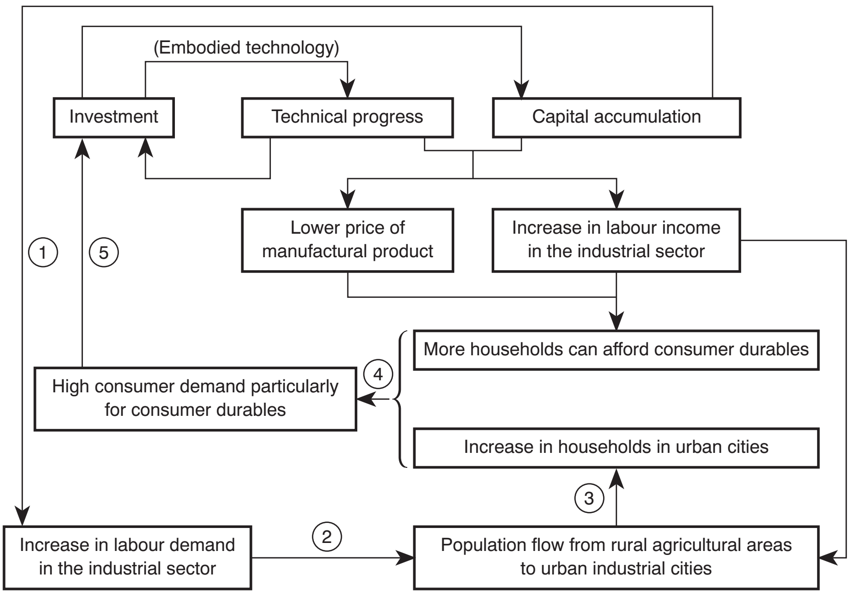


*Figure 2.4* Domestic-demand-led high economic growth of the Japanese economy, 1955–70.

(1931–40) as against 10 per cent in the 1955–70 period. Chenery, Shishido and Watanabe (1962), using the input–output analysis, also drew a similar conclusion on the difference between the pre-war and post-war Japanese economies.

The domestic demand-led virtuous circle for economic growth based on the Lewisian dual structure is not unique to the post-war Japanese economy. Kindleberger (1967, 1989), for example, discusses the post-war growth of the European economy in a Lewisian model; See also Boltho (1982). The relation between household formation and economic growth is also well-known. The so-called building cycle has a long tradition (See, for example, Hickman (1974)). However, in the case of the pre-war US economy, particularly prior to the restrictive immigration law of 1924, it was immigration rather than domestic population flow that endogenously responded to the growth of income and fueled household formation. And in the building cycle, it is residential construction that brings about a close relation between household formation and economic growth.

In the post-war Japanese economy, it was firm's fixed investment rather than residential construction that played a crucial role in growth and fluctuations. In the 1960s, residential construction was less than half of total private construction. Its share rose to above 70 per cent in the mid-1970s when the growth rate sharply fell. Household formation by way of diffusing consumer durables rather than residential construction, ultimately sustained high investment in manufacturing industry. In this sense, my explanation for the Japanese economic growth during the period of 1955–70 is more similar to Gordon (1951) than to the building cycle theory.

*Table 2.3* The relative contribution of demand components to the business cycle (%)

| | *GNP* | *Inventory Investment* | *Housing Investment* | *Fixed Investment* | *Consumption* | *Exports* | *Imports* | *Government Expenditures* |
|---|---|---|---|---|---|---|---|---|
| Japan | 100 | 15.8 | 16.6 | 58.2 | 45.0 | 27.8 | –51.8 | –11.6 |
| US | 100 | 24.3 | 20.8 | 25.1 | 35.5 | 11.5 | –17.5 | 0.4 |

*Source:* Yoshikawa (1993). See the text for details.

Gordon argued that high investment, both residential and nonresidential, in the US in the 1920s was fueled by urbanization and diffusion of the automobile.

In the Japanese economy during 1955–70, population flows and the consequent household formation by way of diffusing newly available consumer durables continuously stimulated economy-wide investment demand. Fluctuations in investment demand were in turn the most important generating force of business cycles from the late 1950s through the 1960s. To demonstrate this point, I show the extent to which different demand components have accounted for different shares of the change in GNP (Table 2.3). Since the Japanese economy has been growing rapidly, almost all variables increase in absolute terms even in recessions I therefore first calculated the change in each variable measured from trough to peak in case of a recovery, and from peak to trough in case of a growth recession. I then subtracted the latter from the former to obtain the difference. Table 2.3 reports the relative contribution of each demand component to this cyclical difference in the change in real GNP. It is the average for the eight cycles during the period beginning February 1957 through March 1990. For the sake of comparison, I also present results for the United States.

In Japan throughout the whole period, the relative contribution of fixed investment has been the greatest of all the demand components: 60 per cent of GNP on average. In contrast, in the United States fixed investment accounts for only 25 per cent on average of the change in real GNP. The relative contribution of inventory and housing investments is greater in the United States than in Japan. Changes in housing investment in Japan are not really systematic over the business cycle. On the other hand, until the mid-1960s, inventory investment had a large impact on the Japanese business cycle: a 60–70 per cent contribution. A substantial portion of the inventory investment was, however, raw materials – which were also imports. Therefore, the contribution of inventory investment and imports almost canceled each other out. As a result, fixed investment retained its importance. As a long-term trend, the role of inventory investment in the business cycle seems to have diminished in both Japan and the United States.

Net exports have been counter-cyclical in Japan's business cycle except for the years 1977–85, in which economic growth was export-led as we will see shortly. In particular, imports have been very counter- cyclical: the fraction of output was –52 per cent on average, compared to –17 per cent in the United States. Until very recently, the bulk of Japanese imports consisted of raw materials and therefore moved very mechanically in parallel with the level of aggregate economic activity.

The contribution of consumption to GNP seems to be in large part similar in the two countries. As for government expenditures, we find them counter-cyclical for Japan (–12 per cent of GNP on average) but neutral (0.4 per cent) for the United States. In sum, the major differences between Japan and the United States lie in the facts that fixed investment plays a much larger role in the business cycle in Japan than in the United States.

Coming back to the high growth era, we find that the situation changed dramatically around 1970. By then the pool of the so-called disguisedly unemployed in the agricultural sector had been largely exhausted. Therefore the population flow from the rural sector and the associated urban household formation both sharply decelerated. At the same time, the then available consumer durables saw saturation in the domestic market. In this way, the domestic demand-led virtuous circle for high economic growth was lost. Judging from Figures 2.2 and 2.3, we note that *this structural change occurred around 1970, a few years in advance of the first oil embargo in 1973* (see also Horie *et al*, 1987).

The structural change is clearly seen for individual industry as well as the macroeconomy. Table 2.4 shows the time series of capacity and investment in the petrochemical industry. By the end of the 1960s, the industry had faced the major turning point. The situation of the industry was typical, not exceptional. The first oil shock hit the Japanese economy which had already seen the structural change

*Table 2.4* Investment in the petrochemical industry, 1956–80

| | *Real investment (Index: 1970=100)* |
|---|---:|
| 1956 | 3.3 |
| 1957 | 9.5 |
| 1958 | 9.9 |
| 1959 | 11.5 |
| 1960 | 15.9 |
| 1961 | 27.2 |
| 1962 | 23.3 |
| 1963 | 25.4 |
| 1964 | 37.3 |
| 1965 | 45.0 |
| 1966 | 30.6 |
| 1967 | 42.5 |
| 1968 | 78.3 |
| 1969 | 81.8 |
| 1970 | 100.0 |
| 1971 | 92.5 |
| 1972 | 55.6 |
| 1973 | 44.1 |
| 1974 | 57.7 |
| 1975 | 65.3 |
| 1976 | 50.0 |
| 1977 | 41.9 |
| 1978 | 24.7 |
| 1979 | 24.3 |
| 1980 | 35.4 |

caused by domestic factors explained above. Before we turn to the post high growth period beginning in the 1970s, I will critically review other explanations for the end of high economic growth.

### *2.2.2 Other explanations for the end of rapid economic growth*

I explained the end of rapid growth during the period of 1955–70 by the domestic structural change of the Japanese economy which had occurred by the end of the 1960s *prior to the first oil embargo in 1973–74*. A popular view, on the other hand, attributes the end of high growth to the first oil shock. There are two slightly different explanations. Note that our problem is to explain a permanent fall in the *growth rate* of the economy, not just a temporary decline in the *level* of output.

Bruno and Sachs (1985) explicitly introduce raw materials and energy into a gross production function. In their framework, the oil shock can be identified with an unanticipated permanent increase in the real price of raw materials/energy. It can be shown that the oil shock is equivalent to a Hicks neutral technical regress. For capital and labour to be fully employed, both real wages and the marginal product of capital must decrease. A decrease in the marginal product of capital becomes larger if for any reason, a decline in real wages did not realize to a full extent. In any case, a decline in the marginal product of capital caused by an increase in the real raw material/energy price entails a decline in Tobin's $q$, and accordingly a decline in the growth rate, *as long as the real interest rate did not proportionately decline*. This is the Bruno/Sachs explanation for deceleration of growth in terms of the oil shock.

I note that in this explanation, the basic reason for growth deceleration lies in the inflexibility of the real interest rate. If the real interest rate declined proportionately with the marginal product of capital, then an increase in the real material/energy price would leave the growth rate intact. Bruno and Sachs (1985) emphasize the role of demand factors in their empirical analysis of worldwide stagflation in the 1970s, but analytically demand and supply sides are not fully integrated. Their analytical framework focuses on the supply side. Yoshikawa (1995; Chapter 4) shows a way to integrate both demand and supply constraints within a unified framework, and also demonstrates that the demand constraints were indeed very important in the case of the Japanese economy in the 1970s.

Another approach which focuses on the supply side of the economy to explain deceleration of growth is Jorgenson (1988). Jorgenson emphasizes the bias of technical change. A 'bias' of technical change indicates how the technical change affects factor shares. Kuroda, Yoshioka and Jorgenson (1984) show that in the post-war Japanese economy (1969–79), the energy using technical progress was dominant in almost all the industries. The energy using technical progress means that an increase in the real energy price makes the rate of technical change decelerate. This is the essence of the Jorgenson argument. The conclusion of his analysis is as follows:

> The most important single factor in the Japanese slowdown is the sharp decline in the rate of technical change. I have now succeeded in linking that decline

> directly to energy prices through the energy using bias of technical change in Japan.

Jorgenson's analysis rests entirely on the traditional growth accounting. His result actually attributes nearly a half of a fall in the rate of economic growth in the 1970s to capital input. And yet, noting that 'after the energy crisis as well as before, the growth rate of capital input was higher than that of output', Jorgenson argues that 'rather than causing the slowdown, the growth of capital after the energy crisis contributed to the continued growth of output at unsustainable levels'. Thus he concludes that 'the decline in the growth rate of capital is not the cause of the slowdown in Japanese economic growth'.

Capital accumulation is certainly not exogenous, however. And what brought about a fall in the rate of economic growth and thereby caused the decline in the growth rate of capital in the first place? Jorgenson's answer is, of course, a fall in the rate of technical change caused by the oil shock. In the growth accounting, technical progress is identified as the part of the blossom of economic growth which cannot be explained by contributions of production factors, the so-called 'residual'. I argue, however, that whether the residual measured by the growth accounting method really captures technical progress or not is an open issue.

Two approaches, Bruno and Sachs (1985) and Jorgenson (1988), both attribute a fall in the rate of economic growth in the 1970s ultimately to the oil shock in 1973. These approaches cannot explain, however, why the second oil shock which occurred in 1979 did not bring about a similar fall in the rate of economic growth; the average growth rates of real GNP for 1973–80 and 1981–90 are 4.1 per cent and 4.2 per cent , respectively. During the first oil crisis, the oil price quadrupled in 1973–74 while in the second oil crisis it only doubled in 1979–80. It might be argued that the first oil crisis hit the Japanese economy harder than the second oil crisis. But transfer payments to OPEC necessitated by an increase in the oil price, when seen as relative to GNP, were actually comparable during the two oil crises, at 3.8 per cent and 4.1 per cent , respectively. The supply side analyses, such as Bruno and Sachs (1985) and Jorgenson (1988), which attribute a fall in the rate of economic growth to the first oil crisis are, therefore, inconsistent with the fact that the second oil crisis did not entail a similar fall in the growth rate. Nor can they explain why the growth rate of the oil-importing Korean economy fell so sharply at the second oil crisis while the effect of the first oil crisis was relatively small, which is converse to the Japanese case. I maintain that demand is an indispensable part of any reasonable explanation of the 1970s.

However, I do not mean to argue that the oil crisis did not affect the supply side of the Japanese economy, but I do argue that a permanent fall in the rate of economic growth beginning in the early 1970s was caused by a *domestic* structural change as explained above rather than the first oil crisis. In this respect, I concur with Maddison (1987). By his careful growth accounting, Maddison finds that the growth rate of real GDP in Japan would have been 3.8 per cent during 1973–84 as against its actual value 3.6 per cent , 'if it had been possible to maintain the relation between energy growth and GDP growth in the previous period' (his Table 15b). In his view, the effect of the oil shocks on growth rate is plainly minor.

## 2.3 The post high growth period: 1970–90

In any case, the Japanese economy saw a major structural change around 1970. Table 2.1 shows that the rate of growth of real GNP was halved from 10 per cent to 4 per cent , but at the same time the Japanese economy became much more stable than in the high growth period. I address this issue of stability first.

### *2.3.1 Investment and the stability of the Japanese economy*

In the 1970s and 1980s the Japanese economy became much more stable than in the high growth period. Taylor (1989), for example, comparing the Japanese and American business cycles notes that 'real GNP fluctuations in Japan are so small compared to those in the United States, especially in the last 12 years that *actual* GNP in Japan behaves much like the smooth *potential* GNP trend for the United States'. Taylor explains the difference between the two economies by taking money as the major shocks to the economy. Nominal wages in Japan are much more flexible than those in the US because in Japan a standard length of wage contract is one year as against three years in the US, and also wage negotiations are synchronized by the nation-wide *Shunto*, or the Spring Offence. Thus, monetary shocks exert less impact on the real economy in Japan than in the US, so goes Taylor's explanation. In the high growth period, however, real GNP fluctuations in Japan were actually *greater* than those in the US. *Shunto* had been in place throughout the period. Taylor's explanation is, therefore, not satisfactory. Then, what brought about the stability of real GNP in Japan beginning in the 1970s?

West (1992), based on a simple macro model of the Japanese economy, decomposes movements of the real GNP during 1975 to 1987 to cost shocks and demand shocks. His conclusion is that demand shocks are estimated to account for nearly nine-tenths of the movement of GNP. This result and my own analysis in Section 2.2 naturally lead us to analyze the behaviour of demand components for 1956–73, and 1974–89. Table 2.5 shows the means and standard deviations of the growth rates of demand components. The stability of consumption measured by standard deviation of growth rate remains the same for the two periods. Exports and residential construction are actually more unstable in 1974–89 than in 1956–73. The stability of real GNP in 1974–89 was clearly brought about by the behaviour of fixed and inventory investment. Here I focus on fixed investment.

There are three closely related causes to explain why investment became much more stable beginning in the 1970s than in the high growth period. First is the effect of aggregation. Investment in the economy as a whole, $I$ is, of course, the sum of investment in many sectors,$I_j$. Therefore the variance of the growth rate of $I$, Var $(\dot{I}/I)$ is the sum of variances of $\dot{I}_j/I_j$ and their covariances. One is apt to forget the effect of changes in covariances.

In the high growth period, economy-wide factors spurred investment and as a result, strong parallels in investment across various industries are observed. This situation changed dramatically around 1970. Yoshikawa (1995) shows that even if the variances of investment in individual industries had not declined, the variance of investment in the economy as a whole would still have declined from 15.6

*Table 2.5* Stability of demand components, 1956–89 (%)

| | *1956–73 Mean* | *s. d.* | *1974–89 Mean* | *s. d.* |
|---|---|---|---|---|
| Real GNP | 9.2 | 2.4 | 3.9 | 1.6 |
| Consumption | 8.7 | 1.7 | 3.6 | 1.7 |
| Private fixed investment | 17.3 | 14.0 | 5.4 | 6.7 |
| Residential construction | 15.3 | 6.3 | 1.7 | 8.4 |
| Inventory investment | 93.6 | 271.5 | 46.9 | 131.0 |
| Government expenditure | 7.1 | 4.0 | 2.5 | 3.7 |
| Exports | 13.2 | 6.7 | 8.6 | 7.6 |
| Imports | 15.0 | 10.5 | 5.2 | 9.1 |

*Source*: EPA, *National Income and Product Accounts.*

to 10.5. At the beginning of the 1970s, investment seems to have been governed by micro-specific rather than economy- wide factors. By way of enlarging the diversity of timing of investment among industries, it enhanced the stability of investment in the economy as a whole. The calculation above suggests that this effect alone explains more than half of a decline in the variance of the growth rate of investment in the second period.

Beyond that, there are two other factors to have enhanced stability of investment. One is an increase in the share of the non-manufacturing sector. Investment in the non-manufacturing sector, compared to that in the manufacturing sector is stable: standard deviations are 14.1 and 8.8 (1968 Q2 – 90 Q4), respectively. An increase in the share of the non-manufacturing sector from 54.1 per cent (1968–72) to 63.8 per cent (1973– 90), therefore, necessarily contributed to the stability of total investment beginning in the 1970s.

The motives for investment also saw a major change. The Development Bank of Japan has surveyed the investment of large corporations by motives. According to this survey, 67.4 per cent of investment was done for the purpose of augmenting capacity in 1969, but its share declined to 28.7 per cent in 1980. Now measured by standard deviation of the growth rate, the volatility of investment motivated by capacity augmentation is much greater than that of other kinds of investment such as labour/energy saving and R and D (research and development): 24.8 per cent and 14.0 per cent , respectively (1977–90). The change in the motives for investment, therefore, also contributed to the stability of total investment.

I emphasize that the three factors mentioned above are all *real* rather than monetary. Some economists argue or used to argue that stability of monetary growth is responsible for the stability of real GNP. Suzuki (1985), for example, argues that

> Since 1975 the money growth rate in Japan has become more stable and the uncertainty associated with it has also decreased due to the announcements of targets. And since 1976 there has been an accompanying decrease in the variability of the inflation and real growth rates.

However, the major influence which brought about the stability of real GNP beginning in the 1970s was investment, and the stability of investment in turn was

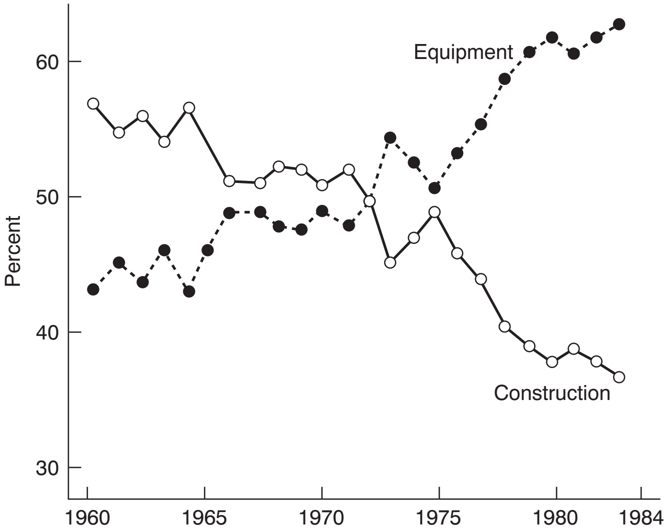


*Figure 2.5* Equipment and investment in plant, 1960–84.

caused by three *real* factors.[1] The stability of monetary growth is basically nothing but a mirror image of the stable real economy.

### *2.3.2 Equipments and structures*

Before leaving investment, it is worth pointing out another interesting fact. Figure 2.5 shows the respective shares of equipment investment and investment in structures. In post-war Japan, during the high growth period the share of structures was higher than that of equipment investment. After the average growth rate fell from 10 per cent to 4 per cent , the share reversed and by the mid 1980s, the share of equipment investment had risen to about two-thirds.

De Long and Summers (1991), on the other hand, using the UN and the Penn World cross-country data, document a strong correlation between the growth rate of real GDP per worker and the equipment investment/GDP ratio. According to their estimates, over 1960–85, a one percentage point increase in the equipment investment/GDP ratio is associated with an increase of about 0.3 percentage points in the growth rate of per capita GDP. They interpret this correlation as causal and argue that it is not just investment but equipment investment that promotes economic growth. The average investment/GDP ratio in Japan is about the same in real terms for both the high growth period (1955–70) and the subsequent low growth period (1970–90). Figure 2.5, therefore, shows that the post-war Japanese experience contradicts the De Long/Summers thesis that equipment investment promotes economic growth more than investment in structures.

De Long and Summers seem to hold that most new technologies are embodied in equipment rather than in structures. However, equipment which embodies new technology may be introduced simply to substitute other inputs such as labour and energy for *given output*. The introduction of new equipment is, therefore, not straightforwardly linked to growth. On the other hand, high economic growth sooner or later, necessitates the augmentation of capacity, and thereby investments in structures. Besides, within a reasonable range, new technology often, if not necessarily, enlarges the optimal size of plant. For Japan, this was indeed the case during the 1950s and 60s. Gigantic steel mills and chemical plants in coastal industrial regions were surely the symbol of the high growth era. They all involved high investments in structures.

All this is, of course, not to say that it is investment in structures that promotes economic growth. Rather the role played by equipment investment and investment in structures in the process of economic growth is much more complex than De Long and Summers assert. Saying that it is equipment investment that promotes growth is too simplistic. In any case, I have already explained how high economic growth was generated during the 1950s and 1960s in Japan. Investment in structures naturally played a major role during the period. On the other hand, in the 1970s and 1980s when the growth rate was much lower than in the previous period, equipment investment was done mainly for the sake of substituting for labour and energy, particularly in the machinery industries and the non-manufacturing sector.

De Long and Summers take the negative correlation between the *relative* price of equipment and the equipment investment share as the major evidence for their argument that high rates of equipment investment were driven by rightward shifts of the supply curve for equipment, rather than by rightward shifts of the demand curve for equipment. However, this argument is not convincing. Since most equipment is tradable, its prices are equalized across countries. When high productivity growth occurs mainly in the tradable sector or the manufacturing industry, the relative price of tradables necessarily declines in those countries where productivity grew: Balassa (1964). This was indeed the case for Japan. The negative correlation between the relative price of equipment and quantities, therefore, does not establish the supply-side causation.

### *2.3.3 The dominant role of exports*

High growth in the 1950s and 1960s was led by domestic demand. In contrast, growth and cycles of the Japanese economy beginning in the early 1970s and up to the mid-1980s were dominated by exports. Vigorous investment in the 1960s had gradually endowed the Japanese manufactured goods with international competitiveness. At the same time, high growth itself also contributed to the competitiveness by allowing firms to take advantage of increasing returns and learning by doing. Given this background, the structural change around 1970 made exports the engine of growth of the Japanese economy in place of domestic demand. This change is clearly visible for the automobile industry, the most important export industry. The growth of the Japanese automobile industry was led by domestic demand in the 1960s but by exports thereafter.

Meanwhile the two oil shocks in the 1970s provided a strong stimulus to the Japanese manufacturing industry. Innovations were particularly vigorous in the machinery industries and by the end of the 1970s the machinery sector came to dominate growth and cycles of the Japanese economy. The first oil shock, though it was an adverse supply shock to the economy as a whole, in fact had some *favourable* effects on machinery industries. In the first place, an increase in costs due to the oil embargo was very slight in machinery industries as compared to other industries such as chemical, paper, and metal; the energy coefficients in the 1973 Input–Output Table were 9.9 (chemical), 7.2 (iron and steel), 4.7 (paper and pulp) and 1.4 (machinery), respectively. Therefore when a sharp increase in the real price of oil, equivalent to terms of trade deterioration in Japan, brought about the 10 per cent depreciation of the yen, machinery industries benefited by gaining international competitiveness. During the period 1973–76, WPI in machinery industries increased only by 25.2 per cent in Japan as against the 40.3 per cent increase in the US. And yet during the same period, the yen depreciated against the dollar by nearly 10 per cent . This means that the price competitiveness of the Japanese machinery industries improved by 25 per cent .

Beyond that, the oil embargo created a huge transfer of money from the oil importing countries to OPEC; about 16 billion dollars in the case of Japan. This newly created oil money was then eventually loaned to developing countries through international financial intermediaries. To sustain domestic investment, growth- oriented developing countries needed to import machinery. In this way, machinery became the product which enjoyed exceptionally high demand, particularly by developing countries, in the generally depressed post first oil-shock world economy. Helped by price competitiveness, the Japanese machinery industries emerged as a chief supplier to meet this worldwide demand.

After the second oil shock in 1979, amid debt crisis many developing countries had to curtail their imports. But in place of them, the US, backed by the overvalued dollar, became the major importer of machineries. In 1970, the share of machineries in Japan’s total exports was still less than half, but by the mid-1980s it had risen to three-quarters.

The emergence of the machinery sector was brought about by vigorous investment in this sector. The machinery block consisting of general machinery, electric machinery, and transportation equipment shares a common catalyst, the integrated circuit (IC). ICs enormously enlarged the potential areas of innovations.

Exports consisting mainly of machinery have indeed become a major force generating growth and cycles of the economy as a whole in the 1970s and the 1980s. For example, the contribution of exports to the growth rate of real GNP was on average –0.2 per cent for the period 1956–70 but it rose to 0.6 per cent for the period 1976–85. For 1980–84, the *relative* (percentage) contribution of exports to growth reached the astonishing level of 40 per cent!

Likewise the simple correlation between industrial output and exports is −0.41 for the period 1960 Q3 to 1971 Q4, but it is 0.56 for the period 1972 Q1 to 1985 Q4. The change in the sign of correlation is consistent with the view that in the 1960s changes in output were not export-led, rather exports were a ‘vent for surplus production’; In contrast, in the 1970s and the early 1980s exports have become a major force generating business cycles in Japan.

*Table 2.6* Variance decompositions of investment, export and gross domestic demand

| | *SE* | *Investment* | *Exports* | *GDD* |
|---|---|---|---|---|
| 1966(IV)–1973(I) | | (10 quarters ahead) | | |
| Investment | 0.068 | 72.93 | 14.48 | 12.58 |
| Export | 0.060 | 29.35 | 39.91 | 30.74 |
| GDD | 0.034 | 51.96 | 20.18 | 27.86 |
| 1973(II)–1984(IV) | | (10 quarters ahead) | | |
| Investment | 0.049 | 43.28 | 51.53 | 5.18 |
| Export | 0.091 | 25.92 | 71.51 | 2.57 |
| GDD | 0.011 | 32.50 | 27.80 | 39.70 |

*Note*: Based on VAR with 4-period lags.

I also estimated a vector autoregression (VAR) model containing quarterly investment, exports, and gross domestic demand to obtain the variance decomposition of each variable (Table 6). Since the variance decompositions stabilize after 10 quarters, the results in Table 2.6 are for 10 quarters ahead. They suggest that investment was very autonomous in the period 1966 Q4 to 1973 Q1 while exports were very autonomous in the period 1973 Q2 to 1984 Q4.

In passing, these results imply first that the assertion that the flexible exchange rate regime is responsible for a fall in the rate of economic growth presumably because it makes exchange rate more unpredictable, is simply inconsistent with the facts, at least in the case of the Japanese economy, since the relative contribution of exports to growth *increased* in the 1970s and 1980s. On the other hand, they are consistent with the fact that the correlation between the Japanese and American business cycles is higher in the 1970s and the 1980s than in the previous period.

## 2.4 The 1990s

Export-led growth created a surge of current account surplus and trade frictions among Japan, US and Europe in the 1980s. At the same time, the yen appreciated from ¥240 per dollar to ¥125 during the short period 1985–87. Yoshikawa (1990) demonstrated that the sharp appreciation of the yen during this period broadly reflected increases in labour productivity in the Japanese export (machinery) industries.

After a brief recession in 1986, the Japanese economy enjoyed an average five per cent growth for four years beginning in 1987. The growth during this period was basically domestic demand-led; It was in sharp contrast to the growth during 1975–85. The biggest problem of this period was that the asset price 'bubbles' accompanied sizable bad investment. When the economy entered recession in 1991, it was left with gigantic bad loans. As of 1997, the amount of bad loans the banks carry is estimated to be 80 trillion yen or 16 per cent of GDP. This is not the place to make premature comments on the current issues, but I suggest that the fragility of the financial sector is *not* the only and the most important cause of the extremely poor performance of the Japanese economy during the 1990s.[2] The average growth during 1991–97 is below 2 per cent .

An important fact is a sharp contrast between manufacturing and non-manufacturing sectors. The Japanese manufacturing industry such as automobile and electric machinery still keeps high competitiveness today whereas the appreciation of the yen made a significant portion of non-manufacturing sectors non competitive. Investment in these declining industries (mostly in non-manufacturing, and particularly in small firms) understandably stagnates, and an extraordinarily depressed investment in these industries is one of the major factors underlying the deep recession in the 1990s.

## 2.5 Concluding remarks

Several important conclusions can be drawn from the analyses of growth and business cycles in post-war Japan. First, the best model to explain high growth of the Japanese economy during the 1950s and 1960s is the Lewis model supplemented by demand factors. Technical progress is certainly a very important factor but the extent to which technical change bears fruit in economic growth depends crucially on how it affects growth of demand; see Aoki and Yoshikawa (2001). Technical change, for example, by lowering the prices of consumer durables, certainly contributed to high growth of domestic demand in the 1950s and 1960s. It also made the Japanese manufacturers a world major exporter of machinery in the 1970s and 1980s. In both cases, technical change raised the growth rate in a roundabout way through its subsequent effects on demand, not through a direct shift of production function. The causality could also run the other way; there is a significant effect of demand on technical progress. The effect of the emergence of the Japanese IC industry on the IC equipment manufacturers is a good example.

Second, exports and the exchange rate are the key variables to understanding the Japanese economy from the mid-1970s up to the present time. Japan imports virtually all its raw materials from abroad, and, therefore, to finance imports, exports have always played an essential role. However, the high growth during the 1950s and 1960s was basically led by domestic demand. In contrast, growth during the period 1975–85 was export led. The appreciation of the yen beginning in 1985 stopped the export-led growth, and, at the same time induced reforms of the declining industries. The sectoral adjustment, however, has not been smooth, and the economy has suffered from a long stagnation during the 1990s.

## Acknowledgement

Section 2.2 draws upon Yoshikawa (1995).

## Notes

1 However, the real business cycle theory cannot adequately explain business cycles in Japan. See chapter 2 of Yoshikawa (1995) for details.

2 Beginning in 1997, the credit crunch exerted substantially negative effects on the real economy, however. See Motonishi and Yoshikawa (1999).

## References

Aoki, M. and H. Yoshikawa, (2001) 'Demand Saturation: Creation and Economic Growth', *Journal of Economic Behavior and Organization* (forthcoming).

Balassa, B., (1964) 'The Purchasing Power Parity Doctrine: A Reappraisal', *Journal of Political Economy* December.

Boltho, A., (1982) 'Growth', in A. Botho (ed.) *The European Economy: Growth and Crisis*, Oxford: Oxford University Press.

Bruno, M. and J. Sachs (1985) *Economics of Worldwide Stagflation*, Cambridge, Mass.: Harvard University Press.

Chenery, H. B., S. Shishido and T. Watanabe (1962) 'The Pattern of Japanese Growth, 1914–54',
em Econometrica, January.

De Long, J. B. and L. Summers (1991) 'Equipment Investment and Economic Growth', *Quarterly Journal of Economics*, May.

Gordon, R. A. (1951) 'Cyclical Experience in the Interwar Period: The Investment Boom of the Twenties', in *Conference on Business Cycles*, New York: National Bureau of Economic Research.

Hickman, B (1974) 'What Became of the Building Cycle?', in P. David and M. Reder (eds), *Nations and Households in Economic Growth*, New York: Academic Press.

Horie, Y., S. Naniwa and S. Ishihara (1987) 'The Changes of Japanese Business Cycles', *Bank of Japan Monetary and Economic Studies*, 5, (3), December.

Inada, K., S. Sekiguchi and Y. Shoda (1992) *The Mechanism of Economic Development*, Oxford: Oxford University Press.

Jorgenson, D. (1988) 'Productivity and Economic Growth in Japan and the United States', *American Economic Review*, May.

Kindleberger, C. (1967) *Europe's Postwar Growth*, Cambridge, Mass.: Harvard University Press.

Kindleberger, C. (1989)'The Iron Law of Wages', in his *Economic Laws and Economic History*, Cambridge: Cambridge University Press.

Kuroda, M. and K. Yoshida and D. Jorgenson (1984) 'Relative Price Changes and Biasis of Technical Change in Japan', *Economic Studies Quarterly*, August.

Lewis, W. Arthur (1954) 'Economic Development with Unlimited Supplies of Labour', *Manchester School of Economic and Social Studies*, 5, May.

Lincoln, E. (1988) *Japan: Facing Economic Maturity*, Washington DC: Brookings Institution.

Maddison, A. (1987) 'Growth and Slowdown in Advanced Capitalist Economies: Techniques of Quantitative Assessment', *Journal of Economic Literature*, June.

Minami, R. (1968) 'The Turning Point in the Japanese Economy', *Quarterly Journal of Economics*, August.

Minami, R. (1970) *The Turning Point of the Japanese Economy* (in Japanese), Tokyo: Sobunsha.

Motonishi, T. and Yoshikawa, H. (1999) 'Casuses of the Long Stagnation of Japan During the 1990s: Financial or Real?' *Journal of Japanese and International Economics*, June.

Nakamura, T. (1981) *The Postwar Japanese Economy: Its Development and Structure*, Tokyo: University of Tokyo Press.

Ohkawa, K. and H. Rosovsky (1973) *Japanese Economic Growth*, Stanford: Stanford University Press.

Shinohara, M. (1961) *Growth and Cycles of the Japanese Economy* (in Japanese), Tokyo: Sobunsha.

Suzuki, Y. (1985) ‘The Macroeconomic Performance of the Japanese Economy and Monetary Policy’, (in Japanese), Bank of Japan, Institute for Monetary and Economic Studies, *Kinyu Kenkyu*, August.

Taylor, J. B. (1989) ‘Differences in Economic Fluctuations in Japan and the US: The Role of Nominal Rigidities’, *Journal of Japanese and International Economies*, June.

West, K. (1992) ‘Source of Cycles in Japan, 1975–1987’, *Journal of Japanese and International Economies*, March.

Yoshikawa, H. (1990) ‘On Equilibrium Yen-Dollar Rate’, *American Economic Review*, June.

Yoshikawa, H. (1993) ‘Monetary Policy and the Real Economy in Japan’, in Kenneth Singleton (ed.) *Japanese Monetary Policy*, Chicago: University of Chicago Press.

Yoshikawa, H. (1995) *Macroeconomics and the Japanese Economy*, Oxford: Oxford University Press.

Yoshikawa, H. and T. Okazaki (1993) ‘Postwar Hyper-Inflation and the Dodge Plan, 1945–50: An Overview’, in J. Teranishi and Y. Kosai eds., *The Japanese Experience of Economic Reforms*, London: Macmillan.

# 3 Productivity–investment fluctuations and structural change

*Bernhard Böhm and Lionello F. Punzo*

## 3.1 Introduction

In this chapter, we review the growth experience in the last two decades of three European countries and compare it with that of Japan and the US. The stylized facts of these economies (Sections 3.5–3.6) exhibit dynamic discontinuities in whose interpretation we employ concepts imported from dynamical systems and bifurcation theories. Relying upon a framework of cross-regime dynamics (introduced in Sections 3.2–3.6) where *regime switches* represent structural changes, our approach departs from traditional and modern growth theories also in two other respects. On one side, we use a sectorally disaggregated description of the economies. On the other, we focus on observed sectoral paths which are typically fluctuating, instead of looking for long run, steady states homogeneous within the dynamic structure of a given economy. In fact, dynamic heterogeneity turns out to be so pervasive that it has surfaced in recent growth literature.[1]

From a methodological point of view, the present exercise goes in the direction of developing qualitative econometrics as suggested by R. H. Day (e.g. in (1992), (1993), (1994)). This approach to modelling stresses emulation of qualitative properties of given historic data sets rather than quantitative estimation and prediction. We make an attempt at marrying it with an inductive framework which focuses upon the interdependence between structure and dynamics. Sections 3.10–3.14 experiment with recent techniques for the econometric modelling of structural change, while the last section reviews some of the open issues in the research agenda on multi-regime dynamics for large scale model economies.

## 3.2 Dynamic approaches

The analytical framework that we employ is constructed as follows. An economy is described as a set of ‘sectors’, just like in certain static models of the input–output family. The relationship with the latter ends there, though. We focus on an economy’s time evolution by trying to reconstruct, from stylized historical facts, the dynamic relationship between capital accumulation and productivity. Employment and its dynamics, which form an increasingly important policy concern in particular for European economies, is left to lie behind the scene, so to say, as a shortcoming of our choice of isolating certain aspects of dynamics in relation to technological change. Thus, our framework and empirical analysis can only give a partial account of the historical evidence. This is certainly a high price to pay for making the first steps in the direction of structural dynamics, the dynamics of interconnected sectors within an economy.

One way to present the heuristic model implicit in our approach is by comparison with the available alternatives which try to explain dynamics through technology and generally conditions of production.[2] For the sake of our argument, one may lump them into two 'traditions' . In the one associated with neo- (and new-) classical theories of growth, modelling style requires that production functions be assumed, factors be measured by their physical quantities, and capital (defined by a variety of measures, as fixed as well as human capital) appears as the stock available at a given date. In other words, stocks are *functionally* related to the flows they are assumed to generate, and the posited production functions relate, at each point of time, productivity, factors endowments and technology. Growth reflects their properties.

We will experiment, instead, with an analytical framework which does not contain production functions explicitly. We maintain that it does not require them, though assuming such functions cannot cause great harm. It may cause some confusion, but it may also help to follow the argument. We try this out for several reasons. First, we share the view that the notion of capital stock becomes rather shaky when we come to make it operational, as it is an estimate whose value depends crucially on a number of assumptions.[3] Second, our empirical analysis is sectorally disaggregated and requires several (sector-specific) production functions. This would amplify statistical and conceptual problems that practitioners, of for example input–output analysis, know very well. Capital stock as an argument in a production function seems to perform better, if at all, in the aggregative parables of growth or long run dynamics, while here, we look for an account or explanation of shorter run sectoral dynamics. The latter exhibits time variability that is hardly compatible with the stability of some underlying technical or production relations. This is all the more important as we allow for the possibility of innovations which, by definition, reflect such instability.

The alternative tradition[4] still offers two possibilities. The 'growth view' sees technical progress as a fundamentally embodied phenomenon and, thus, maintains that investment is its (privileged) vehicle. Productivity growth is faster, the faster the process of capital accumulation. This view claims to be dynamic as it is best formulated in terms of growth rates, rather than levels, of the chosen variables (productivity and investment). At least in one version, it 'regresses' the growth rate of aggregate net output on the growth rate of aggregate investment. The idea of explaining the dynamics of a flow (productivity) with the dynamics of another flow (investment as capital formation, instead of capital stock), thus without using the production function,[5] has received a variety of formulations. Still, the classic one that deploys the 'technical progress function', was first proposed by Kaldor (1957) and later cleared up, analytically, in Kaldor and Mirrlees (1961).[6]

The lack of consideration for stocks and related production functions is a feature typical also of the neo-Schumpeterian view. One can even say that the latter questions the very existence, not merely the relevance, of a functional or causal association between capital stock and investment with productivity. The dynamics of the latter is explained by the pace of innovations, e.g. new forms of organization of production and distribution and/or new products which increase the gap between material input costs and sale prices (i.e. value added). Investment, hence, could really 'explain' neither the time profile of economic performance nor that of technological advance. Models in the neo-Schumpeterian style are thus best formulated as pure flow relations, in terms of variables like employment and productivity measured by value added (though more recent versions do allow for some role for capital). This is reflected in the fact that, in contrast to the two versions of the growth view above, full employment is neither an assumption nor a result. The possibility arises of macroeconomic and sectoral fluctuations which involve the dynamics of labour employment. In some cases, they may be so severe as to lead to dynamic discontinuities, or structural changes.

In synthesis, the analytical framework we are going to present, is based upon only flow variables, which do appear in the latter tradition but are treated differently here, in a context where growth, oscillations and structural change are facets of a unique dynamics. Here too, variables are manipulated so as to obtain growth rates from levels but they are also sectorally disaggregated. The parallel dynamics of the sectors in an economy is thus represented as a set of fluctuations in growth rates, an idea that belongs to the Schumpeterian tradition and was fully articulated by R. M. Goodwin for the dynamics of income distribution in a one-sector, two-class economy.[7] We use growth rates to identify and to characterize *structural cycles*: i.e. a form of basically irregular oscillations across growth paths.

The typical output is illustrated briefly in Sections 3.5–3.6, with reference to the histories of the European economies, Japan and the US.

## 3.3 The Framework Space

We are thinking of an economy as the analogue of a complex dynamical system. It has therefore an architecture given for example by the economy's sectoral wiring. This notion can be naturally extended to a set of interconnected economies. It can also be simplified as when we think of an economy as a single aggregate sector. Hereafter, economies are blocks or systems of interdependent sectors but the logical construction is indifferent to the level of aggregation.[8]

As anticipated, the construction[9] starts with time series of *value added* $(VA)$, *gross physical capital formation* $(I)$, both taken in real term, and *employment* $(E)$. Dividing $VA$ and $I$ by $E$, time series data are converted into intensive form $\nu$ and $i$ respectively. In the following, we will consider the growth rates of value added per person employed and of physical gross capital formation per person employed, as the state variables of the systems investigated. Data sets are taken in a disaggregation into economic sectors, and whenever available the OECD classification of productive sectors is followed.[10] We examine such time series for three European countries (Italy, Germany, France), using Japan and the US as a benchmark.

The *growth rate of value added per person employed in sector j* at a given date (which is omitted here), is defined[11] as

$$g_{v,j} = \mathrm{d}(VA_j/E_j)/\mathrm{d}t*(E_j/VA_j) \tag{3.1}$$

while the *growth rate of gross, physical capital formation per person employed* in sector *j* is

$$g_{i,j} = \mathrm{d}(I_j/E_j)/\ \mathrm{d}t*(E_j/I_j) \tag{3.2}$$

An economy is a set of, say, *k* distinct sectoral sub-systems whose time behaviour is represented by paired time series of growth rates. There will be therefore 2 by *k* pairs at each date in the history of the given economy (from 2 by *k* time series). A pair of values for (3.1)–(3.2) gives the co-ordinates of the dynamic path of the assigned sector in a plane. The co-ordinate space will be called the Framework Space (*FS*) after a certain logical operation has been performed.

The plane is endowed with the *Innovation* and *Accumulation* axes, taken in the conventional order. The former is associated with the growth rate of value added per person employed, the index of productivity growth chosen here. The other axis records the pace of investment plotting values of the variable (3.2), indexed again by both the sector index and the corresponding date. The abscissa axis is the *Innovation* axis as in the neo-Schumpeterian interpretation productivity dynamics is functionally independent of capital accumulation, at least in the extremistic representation proposed before. Conversely, the *Accumulation* axis monitoring the process of change in investment intensity, would be the focus of conventional aggregate theories of growth and technological progress. One novelty in our framework is that the two axes are plotted one against the other.

A sectoral path can now be traced as a sequence of dated states, or pairs of co-ordinate values in *FS*. States are dated according to a ‘clock’, a device defining conventionally the relevant time horizon. One can consider the time intervals used to construct the original data set as the ‘natural’ periodisation. This does not necessarily provide a clock for all investigative purposes.[12] A sort of temporal aggregation, imposing an ‘artificial clock’, can often be useful as a simple smoothing device.

The criteria to carry this out may vary with the phenomena under investigation and one’s own viewpoint. One may, thus, orientate the determination of time intervals on for example the macroeconomic business cycle, measured by the time span between peak or trough values of the GDP growth rate. Alternatively, external or supplementary information can be used to determine a time breakdown based upon the rhythms of some exogenous and/or domestic shocks.[13] These two can be interpreted as macroeconomic clocks and be employed to generate a uniform periodisation across sectoral dynamics. However, a general cycle period may differ sensibly from sectoral cycles and, for certain investigations, the properties of sectoral fluctuations must be retained. A clock and the implied periodisation should, therefore, be considered as experimental devices whose worth depends upon what they allow us to see.

Once the clock has been chosen, a state represents the growth path of a sector at a given date. Thus, in comparison to more conventional approaches, the second novelty introduced here is that a path is observed through two state variables, instead of one. Moreover, overall evolution is reconstructed as a set of growth paths akin to a *segmented trend*[14], exhibiting in general a variety of oscillatory patterns. Of course, segmentation collapses into a single (smooth) trend if we choose the long run periodisation, i.e. a 'date' as long as the total span of the time series available.[15]

The geometric representation in *FS* borrows concepts from the mathematical theory of dynamical systems. Any given sequence of dated states or *trajectory* is, thus, a single instance or realization of the phase portrait in the sector's own phase diagram. As any of the empirically given pairs of growth rates typically jumps dramatically around its state space, sectors are similar to generic dynamical systems in the two-dimensional $(g_v, g_i)$ plane with some complicated dynamics. That the actual history of a sector gives us a single time series of states (single trajectory) implies that it is difficult (if not impossible) to recover the mathematical model, or dynamic equation(s) with appropriate restrictions, which generate such trajectory. However, one such system does exist conceptually and it may take a formulation in continuous or discrete time.

We pick up the latter as it is standard in the econometric and time series approaches. Some of the econometric issues involved in practical model selection are discussed in Sections 3.10–3.14, where also the issue of linear versus non-linear specifications is reviewed. Here, for the sake of the argument, we may simply posit a two-dimensional system of first order, generally non-linear difference equations

$$\mathbf{g}_{j,t+1} = \Phi(\mathbf{g}_{j,t}) \tag{3.3}$$

where vector $g_j = [g_v, g_i]$ is defined on the whole of the *FS*.[16] Although (3.3) may appear structurally simple, it can still be a monster to analyze, depending on the properties of the $\Phi$ maps (recall that $\Phi$ is a 2-dimensional vector). The latter can be simply assumed positing models and family of models parameterized by their associated $\Phi$s. This would be the instinctive attitude of a well trained model builder. If, instead, we refrain from of this procedure and try to recover them from actual data, the issue looks altogether different. It is reasonable to expect that the system be implicitly defined only locally, around an observed path, and then a linear representation might prove good enough, or be all we can get. But then, if we are lucky, we get more than one model (3.3) for a sector. Let us look at the architecture of the economy.

Treating an economy as a system of interdependent sectors, we have two equations in discrete (continuous) time for each of its $k$ component

$$g_{j,t+1} = \Phi_j(g_{1,t}, g_{2,t}, \ldots, g_{k,t}) \tag{3.4}$$

Equation (3.3) is then the special or degenerate case of (3.4) whereby sectoral paths are independent of each other or dynamically de-coupled[17], hence, only 'own' state variables appear. It is suited to represent either a closed aggregate economy or a sector whose dynamics has no functional relation to the dynamics of any other sector. A system (3.4) could be built by simply assembling $k$ such pairs of equations and therefore it would be appropriate to represent a set of economies considered to be close to each other, and/or many parallel but de-coupled sectors. The former is typical of frameworks for cross country studies: if we reproduced the performance of a set of economies in one and the same *FS*, we would be using this formulation implicitly or explicitly. Although such exercises can be performed (and they have been presented elsewhere[18]), in this chapter we focus on the interplay between sectoral paths within an economy, and therefore cannot expect a standard version of (3.4) to have such an orderly block diagonal structure.

On the other hand, we may also simplify a system (3.3) by de-coupling the state variables of the sector. In this case one of the variables would drive the other without feedback from the latter's dynamics. We have a triangular system which can be treated as a one-equation model once we know (or assume to know) the dynamics of the driving variable. Compared to (3.4) this yields a doubly simplified system that can be used for example to illustrate how to begin with an econometric approach (it is used in Section 3.12).

All in all, it is justified to resort to simplified systems either when dynamic independence is an acceptable assumption, or else as a step in the process of learning how to build a model of full interdependence between sectoral paths. Equations like (3.3), for an individual sector are best treated as proper subsystems of (3.4).

## 3.4 Regimes

It was pointed out how the one-sector model (3.3) reveals a twofold difficulty. One relates to mathematical formulation for, if we choose discrete time dynamics as might seem more appropriate, we are already in a realm of chaotic dynamics, given that in principle the system (3.3) is non-linear.[19] The other difficulty is related to the insufficient statistics to identify the explicit form of the system, i.e. the pair of $\Phi_1, \Phi_2$ functions. Both problems are obviously greatly amplified once we allow for a feature of reality, that is the dynamic interdependence among sectors as when positing system (3.4). In fact we are dealing with *k parallel but coupled* two-dimensional dynamical systems (the $k$ sectors). Even if we assume each of them to be relatively simple, the overall or global dynamics can be anything one may imagine. An economy has a *complex dynamical scheme*.

One may hope that, in real economies, interdependences take up some simpler form, with asymmetries introducing hierarchical orders, hence decomposing (3.4) into smaller, in principle computationally simpler, block systems of non-linear equations. At any rate, modelling (3.4) explicitly requires a strategy to choose the $\Phi$s.[20] On one hand, one can introduce a set of hypotheses into an estimated version of (3.4), to the effect of generating dynamic behaviours *compatible* with actual, observed dynamics. The latter approach is based on an understanding of the dynamical system (3.4) as a *reduced form* of some, yet unspecified, model in structural form. Alternatively, in a more standard approach we start with assumptions defining one such structural form, then the corresponding reduced form is derived, and finally simulated dynamics is analyzed in comparison with the actual one. Along this line, it would be natural to begin with production functions, etc. as is typical of production-oriented modelling, and to choose them with the desired properties. The two strategies are, of course, compatible with one another, the difference being essentially methodological. (Section 3.11 illustrates an application of the former and reviews some of the estimation problems involved.[21])

Here we proceed heuristically as follows. If (3.4) above renders the global model of a sector's behaviour, one can segment it into a set of local models giving dynamics under particular conditions, i.e. for certain values of the two co-ordinates. The intuition is that, locally, one can almost always represent dynamic behaviour with a description that in principle is simpler (often, linear) than the overall dynamic model. Such a local model is meant to explain the dynamics from one path to a 'nearby' path of the same family or dynamical class. A *regime* is a family of growth paths that are all generated by one and the same standard model, a *canonical model*, for restricted sets of values of its parameters. Thus, in the *FS* we are going to distinguish six regimes plus one special regime, the Harrodian generalized set. To start from the latter, all paths exhibiting the typically steady state property of time-constant ratios of investment to value added belong to the *Harrodian generalized set*. The latter, of course, includes as special cases those steady state paths of constant levels of capital and output discussed in conventional growth theories (as shown in Section 3.7). The set of paths with investment and value added growing at the same rate (i.e. the 45° line) represents the natural extension. When it comes to empirical analyses, it will be convenient to speak of a *Harrodian corridor* around the line, where the two rates are almost equal to each other, to allow for small deviations, errors and the like, in statistical data. As the problem is general,[22] it is better to keep calling it a set, and recall that it has the typical property of Harrodian paths, being a knife-edge.[23]

The ratios between growth rates in the *FS* yield one of the parameters of the canonical model. They can be either larger or smaller than one, and a ratio of exactly one can be treated as a bifurcation value. It corresponds to all paths belonging to the Harrodian set, and it can be used to characterize that set compared to all others. On the other hand, the four semi-axes can be used to yield the second parameter. All pairs of values of growth rates in the first and third quadrants preserve the same signs (positive value is associated with positive, and vice versa), while for paths in the 2 and 4 quadrants signs are interchanged. This reflects the fact that the underlying relationship between levels of variables, i.e. $\nu$ and $i$, is increasing or decreasing.[24] This can be represented, simplistically, by a second parameter ranging on the real line: for positive (or negative values) we get thus either relationship.[25] The *Harrodian set* together with the other 4 semi-axes in a *FS*[26] can now be used to induce a particular partition into dynamical regimes. Each regime corresponds to a family of realizations of the canonical model for values of the two parameters in partitions of the parameter space induced by their 'bifurcation values', 1 and 0, respectively.

In the *Innovation* regime (regime I) corresponding to the area of the first quadrant *below* the Harrodian set, all paths show positive productivity growth rates exceeding positive investment growth rates. The name was justified in Section 3.2, as in this area the functional association between productivity growth and gross capital formation is nil or weak. Likewise, the area *above* the set, where productivity falls behind investment growth, is the regime that can be associated, though not uniquely, with conventional growth theories which conceive growth as a *capital-driven path* (it is regime VI). With the quadrants numbered clockwise, beginning with the innovation regime – and observing that the positive and the negative quadrants are further subdivided by the Harrodian set – a classification is obtained: with number II being associated with 'restructuring' and showing negative investment growth but positive productivity growth, while the remaining three are mirror images of those just described.

It is only when the $(g_v, g_i)$ co-ordinate plane is endowed with this theory-induced partition, that it makes sense to call it framework space. Traditionally, theories see only regimes I and VI (and their polar cases, III and IV respectively)[27]. The introduction of regimes II and V presents us with the possibility of analyzing oscillations that fall outside standard economic dynamics. Oscillations are now fundamentally across growth paths and these cannot be treated as (sometimes, purely virtual) long run equilibria or steady states.

Dynamics that takes across regimes can be associated with structural change, for it is the 'model of growth' that is changing, then, not just its quantitative properties.

## 3.5 Of country movies

In general, an *economy* may refer to a district, a group of firms or an economic and/or administrative region or finally a macro-region like the Italian North-Centre and the Mezzogiorno.[28] Here, it stands for whole countries: Italy, France, Germany, Japan, the US. Sectors located in a given economy are treated as component systems, location in different economies making them different. The ultimate, binding constraint to the detailed treatment of economies is in the available data, the way in which it is collected and the level of aggregation in which it is finally released. Ideally, systems should be as close as possible to microeconomic decision units. The foregoing results span a set of stylized facts: against them the empirical relevance of certain theoretically defined dynamic behaviours represented in the *FS* may be tested, and assessed also against other theoretical scenarios.

There is a distinct *FS* for each economy. Visualizing its history requires to take as many pictures as there are 'dates', their sequence becomes a *movie*.[29] In each picture, the simultaneous states of several sectors are super-imposed giving a cloud of growth paths. For one sector at a time, a single picture can represent its evolution giving the cloud of paths followed by the sector over different dates.

As changes of regime are structural changes, the dynamics of sectors and/or economies that take them across regimes, is a *structural dynamics*, represented with the level of precision allowed by a low dimensional dynamical model.[30] Our stylized facts refer to such structural dynamics, and do not retain the fine description of the business cycle approach.

We reproduce the movies of five countries (See Appendix 3.2). Any one of the graphs, with all dots erased, is an example of *FS*, showing the partition in six regimes and the Harrodian set. Each movie is a set of five sequential shots or pictures in the respective *FS*, plus one picture that represents the *long run average* path in the two growth rates. The latter is reproduced for comparison with the viewpoint of the conventional growth approach: it looks however at only one of the two rates, the one defined only for the aggregate economy. A uniform clock marks the time for all three European countries; it is derived from the timing of the business cycles in the countries GDP growth rates (peak to peak). This unique clock cannot capture some important features of the observed fluctuations, like the different phasing of Europe with respect to the US and Japan (for the latter, a different periodisation has been used), but this is not the focus of our analysis. In any case, a date is a full cycle; hence, in each picture, the growth dynamics in terms of net rates in the two chosen variables is averaged over the set of years spanning the corresponding business cycle. Likewise, for ease of comparison, a fixed focus has been adopted: in other words, all graphs have the same maximum and minimum values, so that clouds of sectoral paths can be directly compared across time and/or across countries.

We see at once that equilibrium behaviours, i.e. growth paths that persist for longer than one date, *and* paths along the 45° line are generally ‘rare’. Moreover, even when they represent the behaviour of a sector over some date, they are soon abandoned. (This cannot be said to be entirely true of Japan and for this reason it would be better to separate the comments on this country’s dynamics from the others.)[31]

Over the time span of our graphs, Italy, France and Germany share the many difficulties of devising and shaping their participation in the process of European unification. Japan, instead, is exiting the era of fast growth and has to cope with the problems of building up and then adjusting to a new role as an internationally integrated economy. That countries are in quite different initial conditions, is an interesting feature of the exercise.

One can see that, between 1970 and the early 1990s, there is not a single sector in the three European countries selected, which does not cross from path to path. In other words, ‘traverse dynamics’ showing the instability of individual growth paths is a generic property both across sectors and across countries. The actual histories of the sectors, moreover, appear to be made up of sets of traverses taking them from path to path but also from regime to regime. In this light, for the three European countries under investigation, the 1970s and 1980s are years of high instability. Its intensity is not the same, though. Italy seems to be the least stable; France is somewhere in between, at the other extreme lies Germany.

A marked tendency to instability seems to be characteristic of the history of Italy in particular. Although this may also reflect some distortion in the data, we are inclined to believe that it is evidence of a drive towards increasing flexibility in response to increased uncertainty in the economic environment. The Italian literature has seen this in the strategies of the industrial sectors. On the other hand, alternating low and high levels of investment, perhaps, are due to the incapability of facing structural problems (linked with technological lags, geographic dualism, and other issues), and the general inadequacy of the various economic policies implemented.

Italy, with France, had a fast accumulation-driven development in the 1960s and shared the experience of great social unrest at the end of that decade signalling change of political and economic atmosphere. The 1960s were a time of rapid output expansion at low labour cost, and often obsolete technologies, fuelling rapid accumulation. This path came to an end due to causes internal to the countries and to the accumulated effects of the past history, thus totally endogenously. For both countries, a sequence of labour union strikes marked the end of the post-war reconstruction period and the need to find a new development model.

One notices that in the 1970–73 period, France's sectoral behaviours seem to be more dispersed, and along a different direction( the $g_i$-, instead of $g_v$-axis). Italian sectors reacted to the new relative factor prices with a comparatively higher, and sectorally more homogeneous, recourse to innovative behaviours. If development at the very end of the 1960s has seen the restructuring of traditional sectors in the industrial core of the country, with fast productivity gains, this process seems not to have spent all of its impetus yet.

The dynamics of the French economy, on the contrary, appears to be markedly diversified across sectors, though all sectoral paths are in the first quadrant (with a majority of them actually in regime VI of fast accumulation). Such behaviour can perhaps be explained (in accordance with the existing literature) recalling the devaluation of the French franc at the end of 1969, which created sheltered conditions for French firms. Such sheltering effects persist in France while Italy begins to face the need for re-structuring already in the following period, 1973–79. For the French it is the time to invest, with the bonus of generous investment subsidies, in certain industries at least; Italian sectors, making a bare living under the shelter of a French-style exchange rate policy, begin to feel the shortcomings of the relatively high investment levels realized in the earlier period. This induces paths to fall into regime I and slowly move towards regime II, with not a few sectors landing there quite soon. Accommodating macro-policies have the sole effect in both countries of postponing or slowing down a much-needed re-adjustment.

Adjustment comes in the first half of the 1980s. Both countries have to restructure, there is however a process of diversification in the reactions to new environmental conditions between traditional and modern sectors. In Italy, restructuring affects all key sectors of the economy, with dramatic labour shedding (partially absorbed by an expanding tertiary or service sector) and little investment. Here, all evidence indicates that data disaggregated by firm size would be much more informative. Different strategies by big and small firms *vis-à-vis* investment and innovation lie behind the drive for sectoral specialization. The Italian traditional sectors (the stronghold of its SMEs) had shown from the end of the 1960s great capacity to re-organize and to do well in an internationally competitive market. Firm-level evidence points out that this is again the case in these years. With their inventive rearrangements (in industrial districts, network firms, etc.) SMEs are the candidate to play the role of backbone of Italian economy and its rescuer of last resort. In France, the slowing down of the investment pace is somewhat more diversified across sectors.

The years between 1984 and 1988 see in both countries an apparent return to a more or less dominant accumulation-driven growth model. A new phase of accumulation seems to be under way. During the period 1988–92 (or 1993), both economies re-enter a phase of restructuring, more pronounced in France, possibly related with the creation of the common European market.

While Italy, on average, during the 1970s is going through a process of accelerating investment, restructuring and innovation introduction with quickly harvested productivity gains, Germany begins with a process of restructuring especially in manufacturing. This is triggered by labour market rigidities and income distribution dynamics, that is favouring labour share to the detriment of profits. The first oil shock is accommodated better than in the nearby countries (or Japan and the US for that matter) by refuelling the economy with public deficit that sustains the pace of accumulation in certain sectors. It is the next oil shock (in contrast to what happens to the other European countries) that brings about a dramatic change in the dynamical scene, causing a remarkable and widespread shrink in both productivity and the accumulation paces across the sectors (a shrinking of the sectoral cloud that proves to be much more serious than in the other countries). With the beginning of the process of labour shedding, capital formation comes to almost a halt, with some investment being diverted to R and D (at least according to external information available in the literature). This re-allocation of investment outlays from traditional capital goods to R and D and other likewise intangible assets, seems to be a common trend that begins to manifest itself in all European countries, except the UK, and perhaps including Japan, and it is one of the ways (perhaps a diffused one) of realizing industrial restructuring[32]. In the next period the overall sectoral behaviour shows a resumed path of fast accumulation. The predominant dynamics in regime VI last well into the final period of our investigation, i.e. 1988 to 1992, a period in which, on the contrary, the other two European countries already fell into the incumbent crisis of the early 1990s. This phenomenon is largely attributed to the positive external shock of the German re-unification.

The relative predominance of dynamics in the regime of accumulation driven growth which is characteristic of Germany *vis-à-vis* France and Italy, emerges also from the inspection of the graphs for the US. In other words, the pattern of structural fluctuations shows more similarities between Germany and the US, than between the former and the other two European countries. To compare, the US begins the 1970s still on a sectorally widespread path of fast accumulation that continues, though at a lower average speed and at the cost of productivity slow down, almost through to the end of the 1970s. The period across the two decades, comprising the second oil shock, sees a set of measures that, in relative terms, in comparison to the other countries, managed to cushion investment preventing a dramatic and widespread fall. The sectoral cloud spreads out over various regimes in the next and the final period, reflecting indirectly the sectoral reallocation of resources linked with the restructuring of traditionally important production sectors.

## 3.6 Some stylized facts of structural dynamics

From this slice of the European growth experience, one may conclude that the relative weight of innovative behaviours has been on the whole moderate and certainly only cyclical. Search for greater flexibility, which has been said to have generated 'development without accumulation', does not seem to have been the driving force behind adjustments to inflation in the 1970s or to the slumps in activity levels at the beginning of the 1980s. These resulted basically in changes in regime dynamics. The US exhibit a different picture, more centred on accumulation-driven behaviours.

These bold generalizations need certainly to be qualified when it comes to Japan. Its history could be better narrated by what it does not show (as compared to the others), than by what it does show. Japan has a system of seasonal concerted wage bargaining (shunto); a tradition of structural or real, instead of simply fiscal or monetary, policies which is comparable with the experience of Italy and France. It has a tradition of economic dirigism which is also typically French. Finally, it also has the large economic presence of the state. However, in its graphs there are a sufficient number of sectors showing a tendency to equilibrium or persistent dynamical behaviours; some of them, and some of the internationally competitive sectors, stay in the Harrodian corridor (as close as they could be) for long stretches of their histories.

These key stylized facts[33] are reproduced also in the sectoral behaviours. After the high growth period of the 1960s there is a marked and generalized slow down and a decreased correlation of investment paths across sectors. There is a fall in growth rates in the 1980s compared to the exceptional performance of the 1960s. This confirms the thesis that the country entered the 1970s already in the middle of a major structural adjustment, but it managed to surf over the troubled years of the 1970s and 1980s. Therefore, Japan's cross-sectorally dominant behaviours lie in the first quadrant, between the innovation and the accumulation regimes, actually with a relative predominance of the latter. There is a very limited area and a short time for re-structuring via regime II, hence by labour shedding (which is dominant in the European countries, instead). This is experienced only by some sectors, not the prominent nor in particular those classified as traditional ones (an interesting fact when compared to the experience of Italy). On the other hand, the new role of driving forces for the country expansion played by sectors such as general machinery, automobile, and electronics, is clearly prepared by their innovative behaviours in the 1970s. The re-shaping of the structure of the Japanese economy taking place then, is evident from the picture of 1969–72, preceding the first oil shock.

In all countries, sectors exhibit structural cycles but, compared with their European counterparts, for a typical Japanese sector the sequence, the number of regime switches and their timing are generally different. Japan's restructuring takes place in the 1970s (in particular, it is accomplished after the first oil shock), compared to the later adjustments in Europe. Here, the intuition is that Japan has been forced to adjust immediately to the new relative prices for raw materials, searching for new technologies (hence restructuring in the first regime) and realizing substitution investment in the 1980s (at a lower level of macroeconomic performance). Reaction in the European countries has, at first, taken the form of an assorted variety of short term adjustments induced by macro-policies (perhaps in the belief that the shock was only temporary and isolated). The second oil shock forced them to realize that something had changed at a fundamental level. The capability to respond with structural adjustments, was far greater in Japan.

This shows up in a stable structure with a dynamics founded upon investment expansion under balanced conditions shared by all sectors. For this, no unique explanation exists, a mixture of policy considerations, cultural values strongly shared by the country and economic factors all had a role to play.

This sketchy illustration of the empirical use of the *FS* is meant to isolate some stylzsed facts of structural dynamics.[34]

- Structural change has been the all-pervasive phenomenon throughout the period under observation, seeing repeated regime shifts in a large proportions of sectors and of economies investigated.
- Structural change *phasing* was different across countries, a facet which can be partially attributed to political business cycles and other related phenomena (e.g. implementation of specific economic and industrial policies).
- Sectoral structural fluctuations seem to follow different patterns in Europe and the US compared to Japan.
- Innovation and capital accumulation are not alternative growth behaviours: they belong to regimes alternating in irregular sectoral fluctuations.
- Finally, Harrodian paths are rarely observed and generally they are short lived.

The latter observation is quite important for the ensuing discussion: the Solow-path belongs to the same set. But, if we compare our representations of the long-run (the last picture in each movie), they tend to look similar, which seems to support the hypothesis of a steady state prevailing in the long-run. Against our structural dynamics that is working in the medium run, this observation suggests also that long run (equilibrium) theories basically disregard important segments of the histories of actual economies. They rely on the optical deformation created by comparing pictures that are taken at some time distance and leave out the history to which they belong.

It is this history that reveals the interesting feature, it exhibits the fluctuating character of economic dynamics and the changing structures that support it. It can be directly compared with some of the findings of the growth empirics literature. (See, for instance, van Ark and Toniolo (1996).)

## 3.7 Growth theories and the fluctuating dynamics in the *FS*

This was a brief review of what can be seen by applying our analytical framework to data that are normally analyzed by a growth approach. The fact that growth rates are here allowed to fluctuate, shows that the *FS* dynamics tries to embed growth into the more general phenomenon of economic oscillations. With the possibility of moving across regimes (travelling along special 'traverses'[35]) such structural dynamics incorporates business oscillations and growth as special cases.[36]

For a comparison with growth analyses, we shall briefly consider the neoclassical aggregate theory of growth and the disaggregated one associated with the multisectoral model. Both versions employ, explicitly or implicitly, some notion of production function. We have argued that our approach does not (necessarily) depend on production functions. This is in a sense trivially true as, technically speaking, we have no 'formal model'. Or, at least, not a unique model: for formal models we depend on others, and this is what we want to show in this and partly the next sections.

Aggregating over our sectors, we may obtain the growth rates of aggregate productivity and investment intensity. If then, we average them over the whole time period (collapsing the sets of dates into a single one), we get something resembling long run values. Finally, projection onto the horizontal axis of the history of a given economy thus condensed by a single long-run path, gets us the key ingredient in the aggregate growth description. Output is GDP, and hereafter will be denoted as usual by $Y$. It can readily be calculated as the sum of VAs. Theory starts with level values of aggregative variables, $Y, K$ for total capital stock, $E$ and $N$ for total employment and labour force, and so on so forth. From this first level are derived ratios like $y \equiv Y/E, k \equiv K/E$, and the like. Finally, we obtain growth rates to describe dynamic properties of growth paths. In the *FS* also, we follow a similar three-layer procedure, keeping the hierarchy between the lowest layer of raw data and the subsequent layers which are obtained via simple manipulation. In other words, just like in the standard growth approach, from the lower we also derive a unique 'higher level' description. In the opposite direction the procedure, of course, does not lead to unique identification, at most it leads to a whole class. That is, a path in the *FS* is associated with a whole set of pairs of time series in the level values of the variables. (This is an implication of the parameterization introduced above to define regimes.)

Now it can be argued that a notion of regime(s) can be traced in growth theory and is implied by much of recent growth empirics, though there it is interpreted as an equilibrium behaviour. In fact, neo-classical growth theory is said to imply a 'prediction' on the long-run dynamic behaviour of an economy. This says that, under certain conditions (basically reflecting properties of the production function), (i) there will be a unique steady state value of $y$, the output per capita, equal under full employment, labour productivity; and that (ii) the long run growth rate of output per capita obeys the equation: $g - n = \lambda$, where $n \equiv \mathrm{d}N/\mathrm{d}t \;\, N^{-1}$, the natural rate of growth, and $\lambda$ is the rate of growth of technological progress. The long-run rate of growth of per capita output would then be controlled entirely by exogenous technological progress. In the absence of the latter, it would be equal to zero. One can express this result by saying that the endogenous growth rate of per capita output $g_y = (g - n)$, and of labour productivity (if $e = n$), are both equal to zero.[37]

One can write this growth rate as a co-ordinate value on a real line and call the latter the growth line of per capita output or productivity. Each $g_y$ value is associated with a growth path. Hence, in the absence of technological progress of exogenous type, the neo-classical theory says that the long-run steady state value of $g_y$ lies at the origin of this line. If it is not there, this is due to $\lambda$ not being zero.[38] Thus, the productivity growth line decomposes into the union of an equilibrium set, made up of a unique point, and the set of all other paths.[39] While the equilibrium path is sustainable in the long run, all other growth rates can only be associated with short run dynamics. At any rate, stability in the large implies that, eventually, the observed path is the one associated with the origin (or with the exogenously fixed value $g_y = \lambda$). Any other path describes transient dynamics.

Notice that, due to this twofold property, uniqueness of equilibrium and system global stability, this is a non-linear dynamical systems with properties in a sense typical of a linear one.[40] On the other hand, notice also that many dynamic paths that are associated with different time series in levels (e.g. in the values of Y, K, E and N) are zoomed into a single equilibrium path on the $g_y$-axis. This shows that the latter is like a one-dimensional version of our Framework Space. Hence, in principle, we can introduce a two-regime classification here too: the systematic equilibrium behaviour, and the transient dynamics. An analogous idea can be found in, for example, the endogenous growth literature which, re-phrased in our jargon, distinguishes between a Solow-type regime (with diminishing returns to capital) and a 'non-Solow' regime. Any growth rate greater than the Solow rate is sustainable in the long run if constant returns to capital are at work, hence parameterization of the implied model is made with reference to the parameters of the underlying production functions. The introduction of two distinct regimes on the $g_y$-line rationalizes the possibility of settling into different steady states.[41] The related partition can be used to explain as an equilibrium outcome, the persistent difference of growth rates (increasing divergence of growth paths) across countries, a feature of recent growth empirics that has found large support in the evidence.

We can re-set the above argument to show some other of its implications and ramifications. The Neo-classical theory 'exhausts' itself on the $g_y$-axis, for the steady state goes along with full employment and the long run rate is exogenously fixed by demographic forces as well as by technological change. There is no need for supplementary endogenous explanatory variables, except for the adjustment or transitional dynamics. It is really this that, together with the full employment hypothesis, implies uniqueness of the equilibrium solution.[42] This is also the conceptual basis for cross country studies. With each economy one can associate the observed growth path as a pair of values for $y$ and the productivity growth rate. All such paths however, monitored through growth rates, are expected to eventually converge to one and the same value on the growth line. It is the simplicity of the neo-classical model, that the properties of one dynamics mirrors to a large extent the properties of the other. Still, we have a dual dynamics between levels and growth rates that is put together via the various notions of convergence (but, in particular, with $\beta$-convergence).[43]

The key idea that has emerged in the literature, is that there may be more than one long-run growth rate, so that countries need not converge to a unique long-run value of $y$ (nor to a small interval around it). It is to account for the apparent diversity of the growth phenomenon across countries, if not yet to 'explain it', that the single axis of productivity growth is not enough. Kaldorian and (all) other endogenous growth theories do exactly this: they introduce extra axes to explain the long run. They are at variance for their choices of such axes.

Thus, some of the debates raging in the recent, theoretical and empirical, growth literature can be re-interpreted as statements about long-run dynamics in a one-dimensional or a two- (or larger) dimensional version of our *FS*. In the latter, one can enrich the picture on the $g_y$-axis in different ways, e.g. admitting the possibility of multiple steady states. This idea extends naturally to the analysis of many countries, belonging to different clubs and then settling in the neighborhood of a distinct long run attractor. Or else one can do country by sectors studies, like those in our 'movies' of Appendix 3.2. In all cases, stability is an open issue and a sensitivity to initial conditions exhibits properties typical of some non-linear chaotic systems.

Therefore, in our two-dimensional *FS* there is room for all long run theories: at its origin, if they are of the neo-classical sort; at its right, if they are projections of paths in two (or more) variables, one variable being used to explain the other. Hence, exogenous and endogenous explanations of growth can be taken to represent two classes of 'models of long run dynamics' that need not be excluding each other. They may replace one other at different times as well as in different economies or sectors. This depends on the empirical evidence.

However, to capture the stylized facts of the experience of growth, we need to do more and account for the whole *FS*. The common long run view of the growth theories confines itself to the first quadrant and to the equilibrium phenomena that may appear there. To charter the rest of the plane, we need to look at dynamics in the 'shorter run', shorter than the time span preferred by the growth views, and allow for out-of-equilibrium dynamics. This is the dynamics taking across paths.

Treatment of investment as the second variable brings in the instability due to its typically volatile time profile. Growth theory can speak of stable monotonic dynamics as long as investment is kept out of the picture (e.g. by confining it to explain the short run adjustment at business cycle frequencies). Once a shorter time horizon is chosen, investment behaviour re-appears. No surprise, then, in the *FS* the typical mode becomes oscillatory dynamics in growth rates. In those instances where it is fairly regular, it resembles a growth cycle. The empirical evidence shows that in general it is not of such a simple dynamic variety.[44] In the appropriate time horizon, in fact, the interplay between the chosen variables may take up different forms: multiple schemes, time varying schemes. To accommodate this feature, the model implicit in our *FS* has to be a non-linear system of simultaneous equations with (at least) two state variables for each sector.

In this light, the Framework Space is a heuristic tool to classify certain empirically observed 'growth phenomena', when they can be actually observed, and inductively produce a usable theory of actual facts.[45] And it is proposed as a first step for an approach that recognizes the possibility of different theories being consistent to each other.

## 3.8 Some related issues

The *FS* has been applied to economies in a multi-sectoral breakdown. A few words on the tradition of multi-sectoral dynamics may be appropriate. It is useful perhaps to recall something of this approach, that, in fashion not so many years ago, was progressively marginalized from macroeconomics by the advent of models populated by representative agents and likewise representative firms. Multi-sectoral theories are in a sort of nobody's land, for they deploy the notion of a 'sector' which is normally neither the macroeconomy nor the individual firm, though it has some statistical advantages in the way some databases are collected. There, we basically encounter growth theory but virtually no account of oscillations. These models talk of long run dynamics in terms of flows, and focus upon equilibrium paths, steady states in a generalized form. Along a balanced growth path all sectors can grow if they grow at the same rate, and if they do so, the path of the economy enjoys certain optimality properties. This balanced growth path, in fact, is the fastest path an economy can afford in the long run and it is expansionary if (or as long as) there are no stock or resource constraints. Moreover, what is important for our argument is that it is endogenously determined in the sense that it is the result of purely technological properties. The optimal and cross-sector sustainable rate is, in fact, a purely technological index. As a consequence, differences in balanced growth rates across economies (like their changes over time) reflect 'technological change' and possibly labour supply (or other resource) constraints.

Multisectoral modelling focuses on an analysis of growth as a cloud of (sectoral) dynamical paths, with sectoral employment practically undetermined. We borrow both notions into our framework, but we enrich them with a description of 'shorter run dynamics'. This may also account for 'why growth rates differ' across sectors within the same economy or for the same group of sectors across different economies. We also do away with the full employment assumption typical of aggregate growth theories.

Let us now turn to the notion of regime as used here in comparison to its use in the literature. As already pointed out, a notion of regime has found its way into the growth field, though only in connection with the cross country convergence issue. A regime in the sense for example of Durlauf and Johnson (1995) is, intuitively speaking, a steady state value of the relevant variable that represents a local (global) attractor with respect to a (sub)set of countries. Multiple attractors can then be simultaneously present, implying a globally non-linear growth model. Our definition differs for two basic reasons: our regimes are defined with respect to (two) rates of growth and not (the single level of) output per capita. More importantly, they need not be attractors, though obviously this may not be excluded. In our interpretation, a regime is a given class of realizations of what might be called a *canonical model*, a class corresponding to an open subset of the two parameter space. Even though within a class there may be one or more point attractors, dynamics becomes really interesting when it is not so. This turns out to be almost always the case in our graphs. Thus, the local model for a given class may or may not be linear[46]. What is clearly non-linear is the dynamics that takes from one regime to a distinct one. Such shifting from one dynamical model (regime) to another is the two-dimensional analogue of what has been called 'multi-phase dynamics' by R.H. Day (1994). It is related to the Hicksian 'traverse'.

Finally, our definition of structural change is clearly context-defined. Having opted for a description of the economy as a set of 'sectors', any change in their mutual relationships has to be taken to imply a structural change. There is a large literature on this, basically of input/output affiliation, which identifies structure with the set of functional (essentially complementarity) relations among producing sectors (the inter-industry part), with the Final Demand (columns) and Payment (along the rows) Sectors. If this is the notion of structure entertained, any changes in coupling parameters and/or in functional forms linking sectors, is naturally synonymous with structural change.

We have altered in an essential way this static picture of structure. Hence, sectors are not represented simply by flows of inputs and outputs per unit of time; they are rendered in their dynamics through the relative rates of changes of both output (in labour intensive form, hence taking into account dynamics within the Payment Sector) and of investment (again in intensive form, taking thus into account, after the necessary statistical modifications, dynamics within the Final Demand Sector). Therefore, an economy is described by a moving cloud of paths in two variables. Its implicit system law describes a 'dynamical structure', represented by a model of dynamic complementarities, a generalized scheme of dynamical couplings in the sense of R.M. Goodwin.[47] Any change in the qualitative properties of the cloud reveals a change in the underlying dynamical structure, a structural change in this special sense.

The notions of regime and of regime change are introduced to make progress towards identifying such structural changes through a more adequate definition of qualitatively 'similar-and-different' growth paths. Analysis of empirical evidence in the *FS* is preliminary to the inductive identification of such breaks and their causes.

## 3.9 Beyond stylized fact

It is clear that all exercises of the above type identify an agenda for further research. Collecting stylized facts may be a nice exercise, and a large slice of modern growth analysis is concerned with collecting such facts, explaining them by finding regularities, etc. The approach above provides one sectorally disaggregated version of the same sort of growth exercise, and this might be, by itself, a justification good enough for experimenting with it. From this point of view, it shows its close relation with a trend towards more disaggregated analysis surfacing in the recent literature.

There are other uses of our approach beyond reproducing Growth and Business Cycle theories and empirics. One can address issues like 'Is there convergence of the European countries to a common dynamic path, interpreted as a structural oscillation?' and then proceed to discuss the policy implications of the answer to the above question. (This exercise has been proposed in another paper, Punzo (1997).) Or, one may inquire about the relationship between structural dynamics at the sectoral levels and macroeconomic performance. However, beyond facts collection and inductive theorizing, one should make some attempts at explicit model building and treat the *FS* as something more than a descriptive device.

One can try several more or less conventional approaches to model building, that can be best collected into three lines of attack. The first one, naturally, begins postulating a theoretical model. In our case, this should have all the standard ingredients of a growth model which (more than other models in economic dynamics) is typically supply-oriented, and this is just like our framework above. What we would need to do, is to lay down production functions with specific assumptions (as to their arguments and properties) and to opt for exogenous or endogenous growth theories, or even a mix of them. Alternatively, we can employ surrogates of production functions, postulating functional relations between our key variables. In other words, we can adopt one theory and do modelling in the corresponding style, either axiomatically or incorporating hypotheses suggested by previously sifted empirical evidence. From one such model 'in structural form', one derives a reduced form expressing the observable implications of the hypotheses and then proceeds to test them against evidence. This is one version of what we may call the econometric approach. This leads to either accepting or rejecting one specific theory or model, hence in principle rejecting the possibility of having a variety of locally valid models. But, it is on this possibility that the *FS* was built.

An alternative direction of attack tries to model mathematically time series data directly and, by comparison with the former, may thus be called phenomenological. There is a number of formulations along this line, from VAR econometrics to Neural Nets and the like. It is model building along this line that is hinted at in Sections 3.3–3.4, where we introduced some standard dynamic equations that might be able to emulate actual time series (or, at least, some of their properties). This is in fact the development that will be sketched out in Sections 3.10–3.14, where some of those properties are shown to be implied by the dynamic equations, and structural change is interpreted as parameter change.

The final section goes cursorily over the last of these three lines of attack, the one which is based upon the typically non-linear notion of Multiphase Dynamics. There, it is the model that is allowed to change (and to change discontinuously) in the dynamics depicted in the *FS*. Such discontinuous change can be represented with a parameterized model whose parameters exhibit critical values for the associated dynamics. This is the interpretation suggested in Section 3.4 to explain the notion of regime and regime switches.

## 3.10 Econometric approaches to structural change

Structural change in the context of economic growth is here associated with a discontinuous movement between growth regimes. A relationship that is considered to imply a particular growth regime, must therefore change in a specific way in order to be associated with a regime shift. In a descriptive framework one may be able to observe such shifts if growth rates, taken as representative descriptions of the growth phenomenon under consideration, move across certain boundaries which partition the space in association with clearly defined rules. Such rules could relate to specific values or ranges of growth rates or combinations thereof. The latter we chose in our framework, as indicated in Section 3.4.

It must be emphasized that in our framework we depict a system working its way through historical time. What we can hope to construct by using econometric methods is, on the other hand, a model derived from theoretical insights or from data-mining which comes close to capturing the variability of the data and at the same time the implications of economic theory. Several approaches can therefore usefully be applied. In the following overview we intend to show some of the more recent developments in a somewhat simplified manner. It should be clear, however, that even the 'smallest' theoretical model underlying the previous analysis (i.e. the model for an aggregate economy) must be represented by at least two of these econometric relationships – accounting for the symmetric dynamic interplay between investment and productivity – which could be considered as constituting some sort of structural form (in the econometric sense). Nevertheless, we shall discuss econometric issues of structural changes by analyzing single equation models. Generalizations to systems of equations have to deal with the complications arising from simultaneity, co-integration (Johanssen (1988), MacKinnon (1991)) and identification. This can be achieved along the approach adopted e.g. by Hendry, Neale and Srba (1988) and implemented e.g. in PcFiml by Doornik and Hendry (1994).

## 3.11 Traditional econometric models

Traditional modelling may follow the lines of the 'general to specific' strategy of the LSE school (e.g. Hendry, Pagan, Sargan (1983), Hendry (1987)) which concentrates on linear dynamic equations of the form

$$A(L)y_t = B(L)x_t + u(t) \tag{3.5}$$

with $A(L)$ and $B(L)$ denoting lag-polynomials of a specific order, $y$ and $x$ are measured variables of interest and $u(t)$ a white noise error process. In terms of the variables employed in the present analysis, the chosen variables are levels of investment per employed and of labour productivity. The above specification is derived from a general data generation process (DGP) which represents the joint probability distribution of all involved variables.

Let $z_t$ be a vector of observations on all variables in period $t$ and $Z_t = (z_{t-1}, \ldots, z_1)'$. The joint probability of the sample $z_t$ i.e. the DGP is

$$\prod_{\tau=1}^{t} D(z_\tau | Z_{\tau-1}; \Theta) \tag{3.6}$$

with $\Theta$ the vector of unknown parameters. Simplifying this general formulation by adding restrictions will yield an estimable model: The assumptions involved are

- Marginalizing, i.e. dividing the set of variables into a subset of variables of interest and those not of interest $(W_t)$ for the current problem.
- Conditioning by selecting that subset of variables of interest considered to be endogenous $(Y_t)$ which is determined (conditioned) by the remaining variables of interest $(X_t)$. The latter should be at least weakly exogenous i.e. $X_t$ independent of $Y_t$.
- Selecting a functional form.
- Replacing the unknown parameters by estimated numerical values.

Then the distribution can be written

$$D(z_t | Z_{t-1}; \Theta) = F(W_t | Z_t; a) G(Y_t | Y_{t-1}, X_t; b) H(X_t | Y_{t-1}, X_{t-1}; g). \tag{3.7}$$

One can recall some standard definitions, in the present context. Thus, strong exogeneity would require $(X_t | X_{t-1}; g)$, lagged endogenous variables may not influence the exogenous ones. Super exogeneity would in addition require parameter vectors $b$ and $g$ to be independent. Then a change in $b$ would not influence $g$, which makes the specification immune to the Lucas-critique.

The partial log-likelihood function of the model can be written

$$\log L(\Theta) = \sum_t L(\Theta; y_t | x_t, y_{t-1}) \tag{3.8}$$

and forms the basis for estimation. As the chances of arriving at a correct model specification according to the true model and the restrictions found by conditioning and marginalizing are very small, the aim is to find an ‘adequate’ model, i.e. statistically acceptable and not outperformed by rival models.

Assuming that this process has been successful (and in particular, the conditioning step has identified exogenous and endogenous variables), a general dynamic specification of the form (3.5) allows a derivation of long run properties of the relationship between the two variables on the basis of the short run information on the data over the period of observation. Now, to recall our variables, let us interpret $y$ for the level of labour productivity, and re-denominate $x$ with $i = I/E$, the intensity of gross capital accumulation (or gross capital accumulation per employed), and to simplify, let us use national aggregate data. Moreover, for analytical reasons, growth rates in the *FS* will be approximated by logarithmic first differences.

Taking only first order lags in the general equation (3.5), one has

$$\log y = a + b\log i + c\log i_{-1} + d\log y_{-1} \tag{3.9a}$$

which has a static steady state solution for growth rates: $D\log y = D\log i = 0$ given by

$$\log y = a/(1-d) + [(b+c)/(1-d)]\log i. \tag{3.10}$$

If $[(b+c)/(1-d)] = 1$, i.e. with the restriction that $b+c+d=1$, we obtain $\log y = a' + \log i$ corresponding to direct proportionality of $y$ and $i$. If the restriction does not hold we obtain the more general relation $\log y = a' + b'\log i$, where $a' = a/(1-d)$ and $b' = (b+c)/(1-d)$. Thus, the set of solutions of the family of equations parameterized by the unrestricted vector $(a,b,c,d)$, is associated with the origin of the framework space, which is in the 'Harrodian set' because of the specific relationship between the two growth variables.

A straightforward generalization to growth paths other than those in the origin of the *FS* will yield a dynamic long run solution. This is achieved by setting $\log i_{t-1} = \log i_t - g_i$, etc., where $g_i = D\log i$ is a given (and not necessarily zero) long-run growth rate. Then, for any pair of given values of the long run growth rates, we get by substitution the generalized equation

$$\log y = a + b\log i + c(\log i - g_i) + d(\log y - g_y) \tag{3.9b}$$

and finally

$$\log y = a/(1-d) + [(b+c)/(1-d)]\log i - c/(1-d)g_i - d/(1-d)g_y \tag{3.11}$$

showing proportionality of levels as depending on their given long-run growth dynamic performance. We may also interpret this result as a long-run growth rate relationship

$$g_y = (-c/d)g_i + [((b+c)/d)\log i - ((1-d)/d)\log y + a/d] \tag{3.12}$$

which is seen to depend on the levels of the variables involved and shows, therefore, time dependence.

Taking differences of (3.11), assuming the restriction $b+c+d=1$ to hold and the long-run growth paths constant (but different from zero), yields $D\log y = D\log i$, a result compatible with points on the 45° line of the diagram. This is a generalized Harrodian set as it is the linear (one-dimensional) subspace of *FS* spanned by all paths with growth rates for the two variables, equal to each other. A narrow band around it was previously called the 'Harrodian corridor' in view of measurement inaccuracy or, as in the present case, the stochastic nature of the approach. If, on the other hand, the restriction $b+c+d=1$ does not hold, we have a long run steady state of proportional growth off the Harrodian subspace[48].

The value of $[(b+c)/(1-d)]$ will indicate the nature of such a long run state and may classify the system under investigation to which regime the path would belong. (This justifies the treatment of the model as parameterized by that value, as carried out in Section 3.4). If, finally, actual growth paths were variable, shifts of the linear relationships of growth rates would be implied, that depend on second order dynamic movements, the changes of long run growth rates or (long run) growth acceleration. In fact it is easily seen that taking differences of (3.11) is equivalent to assuming a dynamic (first order) relationship in growth rates apart from a constant, i.e.

$$g_y = bg_i + cg_{i-1} + dg_{y-1} \tag{3.13}$$

Its 'static' long run solution (i.e. the one with constant growth paths) is $g_y = [(b+c)/(1-d)]g_i = b'g_i$ as above, while its dynamic long run solution is equivalent to the first difference of equation (3.11). By comparison to this which is the focus of much of growth empirics, paths characterized by variability of growth rates, or 'traverse paths', belong in a sense to a second order dynamics. They seem to be typical of the dynamics in the *FS* associated with any actual economy.

We notice that the short run parameters are compatible with different long run theories expressed by proportionality in levels, growth rates or accelerations[49]. In fact, the rather general, but linear, dynamic model specification introduced above may in principle generate different particular solutions or steady states, depending on whether we assume the given long-run growth paths to be zero, or non-zero constants, or else variable. In order to analyze changes in the structure implying these properties we need to investigate the time stability of its parameters.

## 3.12 Testing for constant parameters

If we remain in the linear framework, testing the dynamic specification proceeds along the conventional econometric arsenal. In view of the discussion about non-linear specification below we may mention specifically the use of the RESET test by Ramsey (1974) which is employed by constructing a more general model including a higher order polynomial to approximate a different functional form. Taking the difference between the general and the (linearly) estimated model one should not find much explanatory power if the null-hypothesis of linearity is correct. So, this test checks the significance of the parameters in the difference of the two models by computing the value of the appropriate $T.R^2$ which is distributed $\chi^2_n$ under the null.

Given the focus of the linear dynamic model on representing steady states it is necessary to devote our attention to the question of stability of its parameters. As the resulting states depend on the values of the parameters, any change in them would also lead to a change in the corresponding steady state. This is especially important for the question of unit elasticity, i.e. the property of belonging to the steady state corridor.

The relationship of the phenomenon of structural change as understood in the Böhm/Punzo approach to the econometric one can be best explained by using a scatter diagram of the two relevant growth rates $g_y$ and $g_i$. The use of prior information about the time periods over which to compute averages of growth rates simplifies the scatter diagram. The points observed now show less fluctuation and usually, due to the selection principle, will show a special pattern that forms the basis for the interpretation of (sectoral and/or regional) histories and their attribution to growth regimes. Discontinuously moving across regimes – i.e. travelling along ‘traverse’ paths – is associated with structural change and it is by visual inspection that changes in regimes are ascertained. The econometric approach typically uses all data of the scatter diagram and fits a curve under the assumption of constant parameters taking account of short run dynamic effects. The implied long-run steady states are those relationships that obtain when short run dynamics has ceased to have an effect. If constant parameters cannot be rejected by parameter instability tests, the model represents one single steady state behaviour derived from the whole period of observation.

Linear dynamic econometric models need, therefore, to be tested for parameter constancy (usually the null-hypothesis) against the alternative of a structural break. The Chow test can be used for this purpose if the time period of the structural break is known in advance, if not then Hansen’s test (Hansen (1992)) informs about potential parameter instability. If the sub-period (for which a different set of parameter values looks likely to prevail) is smaller or equal to the number of parameters one can use the one-step-ahead forecast errors in Hendry’s forecast test which is $\chi^2$ – distributed with degrees of freedom according to the length of the sub-period tested.

An alternative way to temporally locate structural breaks is via recursive estimation which provides a more general framework and does not require prior knowledge of possible breaks. This technique estimates the parameters by least squares over successively increasing periods and, thus, generates a time series of the estimated parameter vector, say $b(t)$ in the model $y_i = b_t x_i + u_i$ for $i = 1, \ldots, t$ and $t = k, \ldots, T$ with $k$ the numbers of explanatory variables and $T$ the total number of observations. The important assumption, however, is that the true parameter vector is constant. We simply derive varying estimates from different data sets. A sudden change in the estimates indicates a structural break, i.e. the relationship has changed its character. Smooth changes are usually indicative of misspecification. Once some type of instability of parameter estimates is discovered the above mentioned diagnostic checks can be applied. Furthermore, the forecast errors from one step ahead predictions can be usefully employed in the construction of stability tests: $v_t = y_t - b'_{t-1}x_t (t = k+1, \ldots, T)$. Under the null hypothesis of constant $b$ and assuming $u \sim N(0, \sigma^2)$, the $v$ are distributed $N(0, \sigma^2 f_t)$ with $f_t = 1 + x'_t(X'_{t-1}X_{t-1})^{-1}x_t$ where $X_{t-1} = (x_1, \ldots, x_{t-1})'$. They are also uncorrelated. The standardized prediction errors are the recursive residuals $w_t = [v_t/(f_t)^{1/2}]$ which are again distributed $N(0, \sigma^2)$. Since they contribute in the updating formula for the residual sum of squares $RSS_t = RSS_{t-1} + w_t^2$, sequential Chow-tests can be constructed, e.g. one step or $n$-step Chow-tests, testing a structural break occurring one or $n$ steps later. They also appear in the CUSUM-test which sums the recursive residuals normalized by the sample standard deviation, and the CUSUMSQ test which basically compares $RSS_t$ with that of the whole sample.

Examples of an empirical investigation using the aforementioned approaches can be found in Böhm (1996). It was shown how the application of careful specification searches and tests leads to identification of regime changes, implying different long-run growth paths.

## 3.13 Discrete switching models

While for the specification strategy mentioned detection of a change in parameters is rather the exception than the rule, because the objective is to find a constant parameter specification satisfying the assumption of linearity inherent in the fundamental model, classes of models specified as non-linear from the outset are able to capture presumed changes or shifts in parameters.

The general class of switching regression models assumes a finite number of regimes, incorporate (observable or non-observable) switching variables and may assume the switch point from one to the other regime as known or unknown (Goldfeld and Quandt, 1972). One can further distinguish cases where the shifts from one to the other regime are discrete or continuous.

As an example of a discrete switching regression model of two variables $y_t$ and $x_t$ take

$$\begin{aligned} y_t &= \alpha_1[1-D(z_t)]+\alpha_2 D(z_t)+\{\beta_1[1-D(z_t)]+\beta_2 D(z_t)\}x_t \\ &\quad +[1-D(z_t)]u_{1t}+D(z_t)u_{2t} \end{aligned}$$

where the errors of the two regimes $(u_{it})$ are identically normal distributed with mean zero and constant variance and $z_t$ is a transition variable. $D(z_t)$ is a heavyside function taking the value of one if $z_t \geq c$ and zero if $z_t < c$, a constant. Since estimation of this discrete model is complicated Goldfeld and Quandt have suggested a continuous approximation to the transition function using the cumulative normal $(c, \sigma^2)$ distribution function with the errors of the regimes assumed to have equal constant variances $\sigma^2$. Using this transition function in the regression will define a smooth transition regression model. This idea has also been followed up in the time series literature (cf. Tong (1990)).

Markov switching times series models may be suitable for time series with jumps or oscillations (i.e. sequences of short jumps). One can proceed along the following steps. First test hypotheses on the number of states present in a particular time series. This determines the number of regimes. Then define the appropriate switching model and its likelihood function. Thereafter estimate the parameters and calculate the ML estimates of transition probabilities.

A simple example of the application of a discrete time series switching model to a two-dimensional problem like productivity and investment intensity growth is the following. Assume that each of the variables may be observed in two regimes, one defined by positive growth rates, the other one by negative ones. A combination of the two variables involved will therefore produce four regimes in which each pair of growth rates can be found. The sequence of regimes obtained can then be read off the simplified time series graph which contains just the averages of values of the whole period in which one regime obtains. Representing this in the phase space gives just four points (i.e. average growth rates), each in one regime, each of which is characterized by one of the four quadrants.

## 3.14 Smooth transition models

A smooth transition between two extreme regimes is sometimes more useful than just assuming the discrete transition between them. As a further alternative to modelling the smooth transition function Maddala (1977) has proposed the logistic function $D(z_t) = (1 + exp(\beta_1 + \beta_2 z_t))^{-1}$. Indeed a number of further alternatives have been proposed and have given rise to a detailed analysis of smooth transition regression (STR) models and have emphasized their attractiveness as compared to conventional switching models.

STR models can easily accommodate more than two regimes, which is more realistic to assume in many cases (e.g. in business cycles, or when considering reactions of economic agents). The two regime model is thus a special case of the STR model. A further advantage of such models is that they can serve as an alternative against which to test parameter constancy in a linear model. Here the alternative to constant parameters is a continuous change in parameters which may be more convenient to handle statistically than a discrete change (as has already been stressed by Goldfeld and Quandt (1972)). Such a STR model can be written by using time as the transition variable: $y_t = (\alpha + \beta D(t))x_t + u_t$. $x_t$ represents a vector of explanatory variables,$\alpha$ and $\beta$ are parameter vectors. If parameter constancy (implying $D(t)$ identically zero) is rejected then $\beta$ can be estimated and the influence of the explanatory variables changes over time according to their marginal effects in $(\alpha + \beta D(t))$. More generally, the transition variable can be any observable economic variable, or even an unobservable one. A specific practical problem is therefore the specification of the transition function.

In a survey of STR models Teräsvirta (1996) presents a selection of popular definitions for the transition function, denoted $G(.)$. Typical are the exponential and the logistic variants. The logistic STR model uses

$$G(\gamma, c, S_t) = (1 + \exp(-\gamma(S_t - c)))^{-1}.$$

$G$ is monotonically increasing in the transition variable $S_t$.$\gamma$, a positive constant (positivity assumed as identifying restriction) indicates how fast the transition from zero to unity is a function of $S_t$. The constant parameter c determines where the transition occurs. The larger the value for $\gamma$ the faster the transition tending to a two regime switching model in the limit. There also exist non-monotonic alternatives to the above mentioned transition function which are useful when reswitching occurs.

Another example is given by the exponential STR model

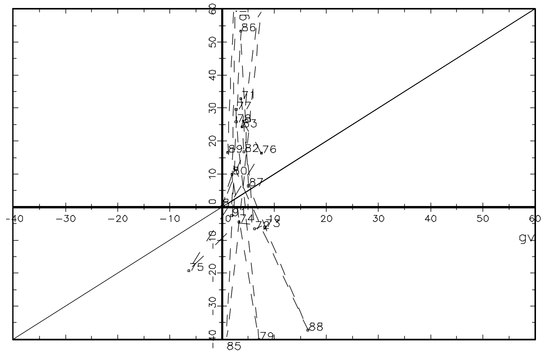


*Figure 3.1* Scatter plot of growth rates of investment and value added per employment

$$G(\gamma, c, S_t) = 1 - \exp\left(-\gamma (S_t - c)^2\right) \text{ with } \gamma > 0.$$

For very large values of $\gamma, G(.)$ will tend to zero and therefore the transitory influence on parameter estimates will tend to vanish. All these transition functions can also employ time as a transition variable and may with suitable changes in their parameterisations also generate nonmonotonic transitions. Examples can be found in Teräsvirta (1996).

The STR models then have the form $y_t = (\alpha + \beta G(\gamma, c, S_t))x_t + u_t$. For specific cases one may choose $S_t = t$. The major advantage of using STR models is that any rejection of the constancy of parameters of the linear relationship is a rejection against a specific parametric alternative. If a rejection of constant parameters occurs, e.g. as typically the result of Chow tests etc., then the parameters of the alternative can be estimated. This helps in determining where and how parameter constancy of the linear model breaks down.

On statistical inference and the model building procedure for STR models consult Teräsvirta (1996). Future research will focus on the applicability of such modelling approaches. A major problem to be expected will be the availability of relatively short annual time series observations which will severely limit the specification search process.

The following demonstration (Figure 3.1) for the paper-production sector of Austria should just serve as example for the potential complexity of the transition function when applied to the investment – productivity relationship. The scatter plot of the growth rates of investment and value added per employment for the sample period shows a dominating variation of the investment rates. Some of them are exceptionally high, especially during those periods when due to environmental legislation most paper producing firms had to engage in investing in new production techniques. In contrast, the productivity growth rates as measured by value added per employment remain with one exception within a narrow band of approximately ten percent. The demonstration thus focuses only on the use of investment as a potentially useful transition variable in the explanation of productivity development.

The dependent variable is the logarithm of value added of the paper sector. It is modelled as a second order autoregressive process with investment and its lag as transition variables in an exponential STR model.

Estimation proceeded by NL least squares. The results of the model

$$\begin{aligned} \log(v34) = a_0 \quad &+ \quad \log(v34_{-1})(a_1(1-\exp(-a_2(\log(i34)-a_3)^2)))+ \\ &+ \quad \log(v34_{-2})(a_4(1-\exp(-a_5(\log(i34_{-1})-a_6)^2))) \end{aligned}$$

are contained in Table 3.1 (using an annual sample 1972 to 1991).

The two estimated transition functions can be inspected in Figures 3.2 and 3.3.

Although the estimation result may still be improved by further specification search, it is easily recognizable that the interaction between the two variables is seemingly complex. Now we need to remember that the dynamic system we are interested in is, in fact, of higher dimensions. Requirements concerning the length of observation periods as well as the task of identifying systems relationships make an extension of these econometric techniques into dynamical systems not an easy task.

A simple simulation study of a two dimensional system with an exogenous transition function may serve as an instructive example. We can show the nature of the simulated paths represented in the framework space to be quite similar to those that have been observed from actual economies.

Let the two endogenous variables be given by

$$y1(t) = (a + b.\,G1(\gamma, c, S_t))x(t) + u_t$$

$$y2(t) = (f + g.\,G2(\gamma, c^*, S_t))x(t) + v_t$$

where $G1$ and $G2$ are two transition functions, $S_t = t$ the transition variable, $x(t)$ an exogenous variable (here a noisy trend), and $u$ and $v$ normally distributed error terms. The transition functions have been chosen as to produce structural shifts during the first third of the thirty observations generated. They are given by:

$G1(\gamma, c, S_t) = 1 - \exp(-\gamma(S(t)-c)^2)$ with $\gamma = 3$ and $c = 10$,
$G2(\gamma, c^*, S_t) = (1 + \exp(-\gamma(S(t)-c_1)(S(t)-c_2)(S(t)-c_3)))^{-1}$
with $c^* = (c_1, c_2, c_3) = (2.5, 5, 10)$ and $\gamma = 3$.

Thus, changes should occur during the second and third and around the fifth and the tenth period. We can graph these generated series $y1$ and $y2$ in diagram (Figure 3.4) and draw a scatter plot (Figure 3.5):

One can observe that there are structural changes in both series around period 10. Whereas series $y1$ exhibits the break only during one period the regime shift in series y2 lasts longer (between periods 2 and 3, and 5, and 10). Turning to their growth rates and the framework space as given by the following graphs in Figures 3.6 and 3.7, one can readily recognize the regime shifts despite the added noise to the two equations in levels.

*Table 3.1* Estimation of NL least squares

| *Variable* | *Coefficient* | *Std.Error* | *t-value* | *t-prob* | *PartR2* |
|---|---|---|---|---|---|
| $a_0$ | 0.41354 | 0.21293 | 1.942 | 0.0741 | 0.2249 |
| $a_1$ | 0.88103 | 0.067007 | 13.148 | 0.0000 | 0.9301 |
| $a_2$ | 20.124 | 59.592 | 0.338 | 0.7410 | 0.0087 |
| $a_3$ | 1.0855 | 0.24027 | 4.518 | 0.0006 | 0.6109 |
| $a_4$ | 0.020500 | 0.010746 | 1.908 | 0.0788 | 0.2187 |
| $a_5$ | 3.1438 | 5.1151 | 0.615 | 0.5494 | 0.0282 |
| $a_6$ | 1.2750 | 0.38476 | 3.314 | 0.0056 | 0.4579 |

R2 = 0.982549 F(6, 13) = 121.99 [0.0000] $\sigma$ = 0.0368252 DW = 2.39
RSS = 0.01762927324 for 7 variables and 20 observations.

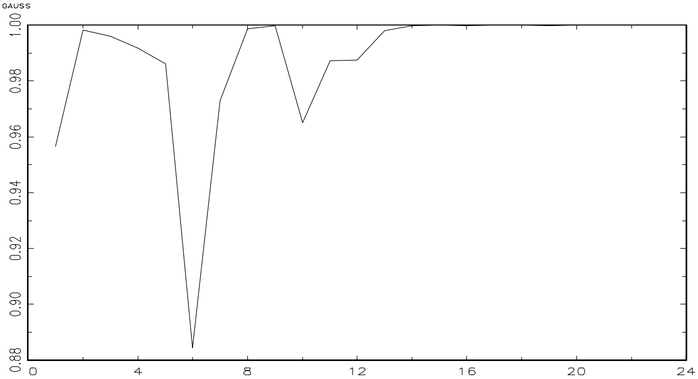


*Figure 3.2* Transition function 1

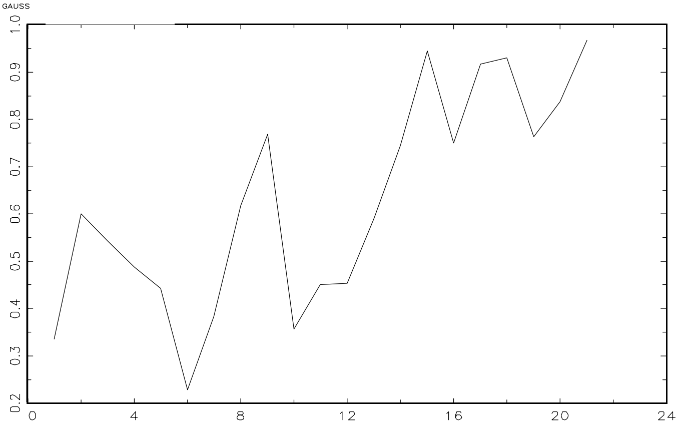


*Figure 3.3* Transition function 2

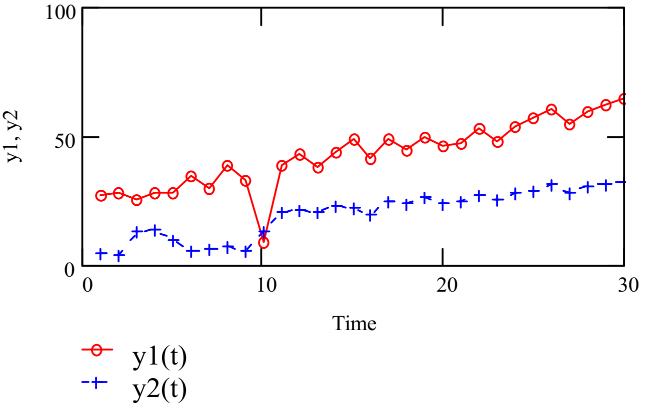


*Figure 3.4* Time series graphs of simulated series

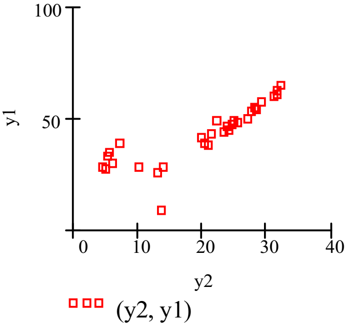


*Figure 3.5* Scatter plot of series in levels

The framework space makes these shifts even more obvious. We find four pairs of growth rates to be located outside the area around the zero point, corresponding to regime shifts according to the construction of the transition function.

The cluster around zero follows from the relationship of both endogenous variables with the exogenous variable, a stochastic trend, which enters linearly whenever the effect of the transition function is zero. Transformed into growth rates both variables have to fluctuate around zero due to the noise component.

We may conclude this excursion into some of the more promising econometric approaches which might shed some light on the possibility to identify relevant economic processes with a quotation by T. Haavelmo in his Nobel lecture (1989):

> I believe that econometrics can be useful. . . . the possibility of extracting information from observations of the world we live in, depends on good economic theory. Econometrics has to be founded on theories that describe in a reasonably accurate way the fashion in which the observed world has operated in the past.

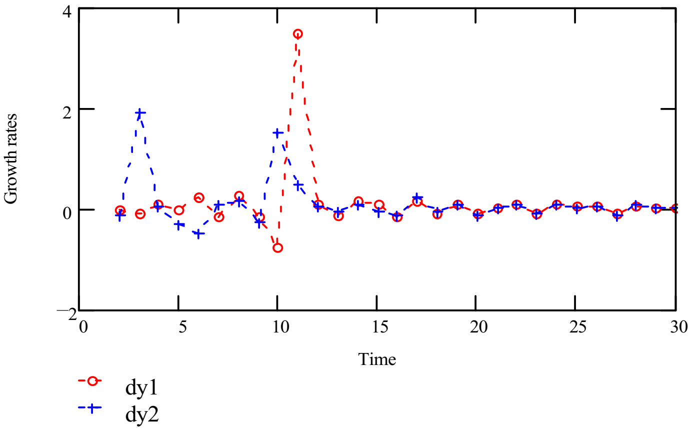


*Figure 3.6* Time series of growth rates of $y1$ and $y2$

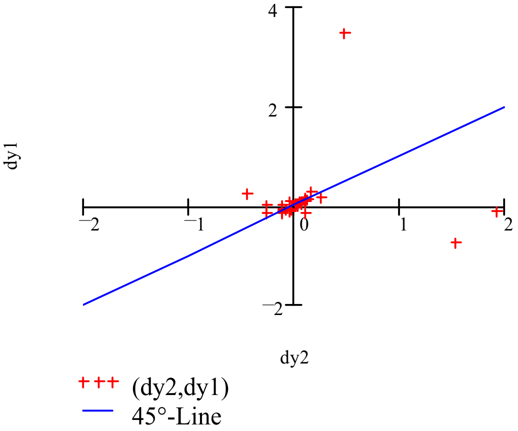


*Figure 3.7* The Framework Space

We have seen that our framework space can accommodate several theories. However, it still remains a formidable task for theories to provide a convincing explanation of economic change and for the econometrician to account for the changing relevance of those theoretical explanations.

## 3.15 Multiregime dynamics in large model economies

As a complement to the econometric approach, rather than in contrast to it, we can do some more mining of stylized facts, and then try out different modelling techniques. Before outlining these further developments, we have to re-consider the dynamics we have been talking about.

So far, we have developed the Framework Space (*FS*) with its structure. This is a partition which is based upon the notion of regime as a collection of states (-paths) qualitatively similar though quantitatively different. Our chosen partition depends upon a classification criterion which is derived from existing theories of growth and cycle.

That each state in our *FS* is a growth path monitored on the basis of two co-ordinates makes for the first of such differences. Furthermore, a particular state can be either a long-run trend or steady state, or else a kind of 'short-run trend' (short as the length of the averaging period of time) when, for example, it belongs to a sequence of short run paths. Thus, in our dynamics there is no role for the distinction between growth and oscillations. Actually, our choice of growth rates as state variables instead of levels, is intended to unify the treatment of the two, so that we start with growth cycles (or cycles in growth rates) as the simplest dynamical phenomenon to expect[50]. A sequence of short run paths may turn out to be regular or irregular to various degrees, both in theory and in empirical evidence. This introduces the possibility (and the likelihood) of encountering chaotic and generally complex dynamical behaviour in our framework. Those regular behaviours studied by a conventional approach[51] are naturally viewed as simple and limiting cases of an otherwise, generically, irregular dynamics of sectorally disaggregated economies.[52]

All this can be expressed by saying that there are three dynamics going on in the *FS*: the first is implicit in the definition of a state as a path; the second takes across paths, the last one is in principle across regimes (that one stays for sometime within a given regime cannot be excluded on first principles). Each kind of dynamics is in a sense typical. The first dynamical layer implies a relation between growth rates and behind it a parameterized relationship between levels (VA, E, I). It can be associated with equilibrium growth theories, which look, among all paths in growth rates, for the one that is a long run attractor ('the data implied steady state'). Recent growth literature on endogenous growth and on non-ergodic growth has expanded the scope of such an approach by allowing for multiple steady states equilibria. At any rate, the growth approach can be said to imply a partition of the *FS* between the subset of attractors and all other dynamical paths that are short run or disturbed paths or 'disequilibria'. The second layer of dynamics is best interpreted in terms of Hicksian traverses: each path is in principle unstable, and the system 'traverses' from one to the other in disequilibrium. What is new is the third one, i.e. the dynamics of structural change according to our definition, where disequilibrium may be systematic.

This refers fundamentally to regime behaviour. Of this it emphasizes the 'qualitative part', in the sense that the dynamical behaviour of the selected system is now looked at through the sequence of the regimes it goes through (or 'stays in'), rather than through the sequence of states. We loose the (quantitative) precision of the state representation, we no longer retain information on the history of where the system was, nor can we try to 'predict' 'where' exactly it will go; these are the effects of the particular texture of the state space when co-ordinates are reals. But we may hope to find out the region where it was/is/or it will go. The information conveyed by the regime description can be better characterized by saying that it tells us where the system does not lie. With this spatial notion, we are practically very close to the point of view of recent chaotic dynamics literature, which is concerned with such problems as: is there a 'region attractor'; what is the likelihood for a given system to visit a particular region of the state space; what is the likely length of time spent in a particular region, and the like.

There is a relationship between these three types of dynamics but one goes from one to the other, in a sense, by a process of reducing information retained. Therefore, there are costs in choosing one or the other, and one must be aware of this. In the following, it is assumed that type-one and type-two dynamics are well known, and we consider only the implications and the possibilities linked to the study of type-three dynamics.

Using the notion of regime implies a coarse partition of our two-dimensional state space which is 'spatially' discretized. Thus, we do not need to use pairs of real numbers as co-ordinates, it is sufficient to use single naturals (and a finite number of them) or else a finite set of symbols in a given alphabet. Any other trick will do. A colouring scheme, like the one often used by Ralph Abraham, could also be used, and would be better for purposes of visualization and animation. In our case, a partition into six regions-regimes is used, but the principle is the same, whenever we can use the notion of regime.

In more abstract terms, we may speak of **k** regimes. Hence we need **k** symbols only (and not two real lines), say letters from A to F. The history of a chosen system can now be encoded through this alphabet, and will result in a string of letters or 'word'. With such dynamic representation, one can address questions like: is there regular regime behaviour? is 'irregular' behaviour dominant throughout the history of the chosen system (or across the histories of a set of systems)? But obviously the terms themselves will need a new definition. To extract such information, if it is there to be found, one can imitate standard econometric techniques, looking across data: (i) within the individual strings that record histories, one string at a time, looking for regular/irregular behaviours; (ii) across the screen of all strings of the sectors within a given economy, to find cross-strings regularities; (iii) finally, country screens can be compared. To characterize the dynamic properties in a string one can resort to ideas that belong to information theory, and in particular to the theory of information complexity.

The presence of a 'fairly' regular 'string' would be the basis of the qualitative approach to econometrics: it tells us of the presence of a stable sequence of regimes that the system may go through. The idea is not entirely new, as something similar was originally formulated in the theory of stages by development theorists. This shows, once again, that in the dynamics of regimes there is no longer a distinction between 'growth' and 'development', the former implying structural stability, the latter allowing for structural change. It is a means to unify them into a unique dynamical theory. This new formalism may also provide the way (or one of the ways) of treating growth and oscillations as disequilibria.[53]

Turning now to the other line of development. Its roots lie in distributional dynamics as implemented by Quah (1993). Quah who is working in the area of 'growth empirics' questions the relevance (and/or interest) of representing the time behaviour of cross-country performance by synthetic properties, like average growth rates, and/or levels of per capita productivity or income. Distributional dynamics takes the original cross country distribution at say time 0 and follows its evolution over time, trying to figure out, from stylized facts, guessing or theoretical construction, asymptotic properties of the distribution itself while keeping track of the intra-distribution dynamics (poor becoming richer and the like). In a discretized version of this (where an income indicator is discretized in intervals of values and countries are classified accordingly), we obtain a dynamics that is easily converted into our coding approach.[54]

At each date, the economy is represented by a state vector with **k** entries, each entry being the relative proportion of sectors in a given economy that is lying in the corresponding j-th regime at that date (that, as usual, can be a whole time interval). The dynamics is therefore represented by a sequence of dated vectors $[x_T]$ telling us the time evolution of the relative frequencies in which each regime is being visited. The sample can be taken to be the sectors of a single country, one sector across a pool of countries, or finally sectors in all countries for which data is available. Between two subsequent vectors one can compute transition probability matrices, and then see if certain tendencies can be identified.

The candidate model in any case must account for a dynamics that takes a system across regimes. This model will have to take regimes seriously, but this is ongoing research.

## Acknowledgement

The author acknowledges the financial support of the Italian CNR, contract n° 98.03803.CT10, and of Siena University through its PAR 1999. Part of the work reported in this chapter is also the result of a close collaboration with other researchers in the *IDEE* Project, financed by the EU within the TSER action of the 5th S & T Framework Programme.

## Notes

1 See for instance Bernard and Jones (1996), and the debate on convergence reviewed e.g. in Durlauf and Quah (1998).

2 Thus, we leave aside models explaining dynamics on the basis of demand behaviour (e.g. Keynesian models) or which conjugate demand and supply conditions in a theory of structural dynamics (e.g. see the work of Morishima , Pasinetti, for instance). We also leave out the neo-Austrian approach, see Amendola and Gaffard (1988), (1998). Our framework is compared with the latter in Amendola, Gaffard and Punzo (1999).

3 For a discussion, see Scott (1989).

4 We mean alternative to the neo-classical one.

5 Again, that there is no production function does not mean that one cannot insert one compatible with the Kaldorian model.

6 There is, of course, a distinct formulation of the model of growth as reflecting embodied technological progress, the one proposed by Arrow with capital stocks of different vintages, but this version is not the one that interests us here.

7 Goodwin (1967); see also Goodwin and Punzo (1987); Goodwin (1990). Punzo (1995) discusses the relation between the present empirical framework and those works.

8 This construction was introduced in Böhm and Punzo (1994) and discussed in Punzo (1995).

9 In fact, one can build an analogue framework with different sets of variables and their growth rates (if this is suggested by the problem to investigate). Notice that gross capital formation could have been re-defined so as to measure capital in a broader sense. Here, we take physical capital in the Kaldorian and Solow type of tradition.

10 See Appendix 2 on statistical data bases.

11 In the econometric sections below, it will be useful to resort to a log approximation of (3.1) and (3.2).

12 Clearly, there is nothing natural about this choice that is dictated by statistical difficulties rather than the timing of the phenomenon under observation.

13 The clock chosen in Böhm and Punzo (1994) reflects the hitting of exogenous shocks, and therefore reports external information. It was used to check, to a certain extent at least, the relevance of aggregate shocks in explaining the structural dynamics of the economies involved.

14 There is some relation with the econometric literature on broken trends here. The idea of a variety of patterns of growth, on the other hand, has recently become popular even in the tradition of endogenous and exogenous aggregative theories. There, however, it is attributed to cross country, rather than to cross sectoral dynamics.

15 This is the choice of growth theory as is done in the conventional mode. Thus, the latter too has a place in the *FS*, justifying the latter's name.

16 The system (3.3) is taken to have no forcing term, an analogue of a VAR formulation. The forcing term may be used to introduce consideration of the influence on system dynamics of an exogenous force, that can have stochastic or deterministic nature or can sum up the two. The forcing term will therefore determine, together with endogenous properties, the system's long run dynamics. For such formulation see again Section 3.11.

17 A notion introduced by R. M. Goodwin in his (1947) contribution, and later investigated in Goodwin and Punzo (1987).

18 Again, see Böhm and Punzo (1994, 1995). The literature on Growth Empirics is full of exercises of the first type.

19 The alternative is obviously to assume that the model be linear, as done in much econometric practice, again see Sections 3.11–3.12, but this has a cost that may be quite high.

20 The dynamic specification of (3.4) as a first order autoregressive process is only used for simplicity. The appropriate order of these processes needs to be derived from econometric testing.

21 Of course, estimation problems enter the latter modelling strategy as well, though at a different stage.

22 It will be the case whenever we will deal with borderlines between regimes in the partition of the state space below.

23 It is a linear one-dimensional subspace, whence the knife edge property descends.

24 Compare with the discussion in Section 3.11 below, in particular equation(s) 3.10, which contains two parameters.

25 And of course, its zero would be again a bifurcation value and correspond to the origin. In this light, the origin corresponds to a bifurcation value of a family of maps with two parameters.

26 We are using orthogonal co-ordinates.

27 Actually the former, i.e. the first quadrant , can be called the growth and the latter, the third quadrant, the contraction quadrants.

28 As in Böhm and Punzo (1995).

29 Movie refers also to the possibility of animating the graphs, a technique amply demonstrated in the work of R. Abraham and C.D. Shaw (1989).

30 It is to be understood that the decision of defining the *FS* in terms of two coordinate variables (the growth rates) already makes the model more complex than a standard growth model. The latter can be argued to imply a one-variable version of our *FS*. Two however is surely less than the desired dimension for describing an economy; however, it already shows all the complexity that a higher dimensional model would have to deal with. The choice is made, among other reasons, to keep a multisectoral model manageable.

31 This and the following sections are based upon Punzo (1997).

32 See *Industrial Policies in OECD countries, Annual Review 1990*, OECD Paris, 1990.

33 See for example Komiya, Okuno and Suzumura (1998), Moriguchi (1991), (1995), Yoshikawa (1995).

34 For more details, the interested reader is invited to look at the publications from the ongoing *IDEE* project, as well as previous publications by the authors and other associates to the project. See also Amendola and Gaffard (1988), (1998); Amendola, Gaffard and Punzo (1999); Böhm (1996).

35 ‘Traverse’ in the sense of Hicks (1973).

36 This belongs to the tradition of Schumpeter and Goodwin.

37 Let $\mathrm{d}\ \mathrm{Log}\ E/\mathrm{d}t \equiv e$ , $E$ being aggregate employment. It is typical of neo-classical and endogenous growth models (including the Kaldorian and Cambridge tradition) to introduce as an assumption a full employment equilibrium path. This is not accepted into our framework.

38 Given a known value of $\lambda$, we can always measure the economy’s growth rate in deviation. Thence, the notion of ‘endogenous growth rate’ is in this case the residual.

39 The latter is the union of two open intervals, the closure being the equilibrium set (a point or one-dimensional manifold).

40 In particular, the long-run state, an attractor, is a monotonic equilibrium path, the simplest of all dynamic morphologies.

41 An idea ingenuously exploited by, for example, Durlauf and Johnson (1995), though in a different analytical setting.

42 If employment levels were let free, in other words, we would have an interval of long-run values. Typically, in growth theories this possibility is associated with short-run dynamics, as they all descend one way or another from Harrodian dynamics.

43 $\beta$-convergence is an exercise whereby a cross country regression is run of growth rate on initial level of output per capita. Hence, a one dimensional *FS* space with a single $g_y$-axis is completed with a $y$-axis. We talk of $\beta$-convergence when the regression coefficient is negative, implying that the poor countries run faster towards the long run equilibrium path. See the vast literature, e.g. Baumol, Nelson and Wolff (1994), and Durlauf and Quah (1998). The resulting framework is not a two-dimensional *FS* in our sense, for one variable is interpreted as the explanation of the other. Econometrically it is a one equation model, rather than the two equation system implicit in our *FS* and described by (3.3) or (3.4) above.

44 See for instance Goodwin (1991), or two collections, Benhabib (1992) and Day and Ping Chen (1993) for a sample of a burgeoning theoretical literature.

45 Changing the variables on the $y$-axis then can lead to different frameworks where the classification between exogenous and endogenous changes too, as they are relative to the explanatory variables on such axis. In our framework where the system is generally simultaneous, the distinction in principle does not make sense.

46 See for instance some of the models used later to discuss the econometrics of structural change.

47 See Goodwin (1990).

48 See the treatment of the model as parameterized by $b' = (b+c)/(1-d)$, in Section 3.4.

49 This result obtained from using a first order lag-polynomial is easily generalized to higher orders. The relevant restriction for long run unit elasticities will then be $B(L)[A(L)]^{-1} = 1$ in the notation of eqn. (3.5).

50 Thus, the emergence of a long run steady state in growth rates is a very special case of a growth cycle collapsing to a single state as a global attractor.

51 We are referring to non-stochastic versions of Growth and BC theories with which our approach can be directly compared. In fact, the latter can be best understood if formulated with a deterministic, i.e. non stochastic, model, where irregularity is endogenously generated rather than shock induced. This does not imply that stochastic phenomena cannot be added to it. Most of modern economic dynamics is formulated in a stochastic environment, to derive some statistically average long-run properties.

52 Actually there is a growing literature on this, (see for instance some of the papers in the collection edited by Benhabib (1992)), though such literature focuses upon the theoretical possibility of observing highly irregular dynamics associated with relatively simple multisectoral models.

53 See Brida, Puchet and Punzo (2000).

54 See Lavezzi (2000).

## References

Abraham, R. and C. D. Shaw (1989), *Dynamics. The Geometry of Behavior*, Aerial Press, Santa Cruz, CA.

Amendola, M. and J.-L. Gaffard (1988), *The Innovative Choice*, Oxford University Press, Oxford.

Amendola, M. and J.-L. Gaffard (1998), *Out of Equilibrium*, Clarendon Press, Oxford.

Amendola, M., J.-L. Gaffard, and L. F. Punzo (1999), ‘Neo-Austrian Processes’, *Indian Journal of Applied Economics*. 8(2) 277-93.

van Ark, B. and N. Crafts (1996), *Quantitative Aspects of Post-war European Economic Growth*, Cambridge University Press, Cambridge.

Baumol, W. J., R. R. Nelson and E. N. Wolff (eds) (1994), *Convergence of Productivity*, Oxford University Press, New York.

Benhabib J. (ed.), (1992), *Cycles and Chaos in Economic Equilibrium*, Princeton University Press, Princeton.

Bernard, A. B. and Jones, C. I. (1996) ‘Comparing Apples to Oranges: Productivity Convergence and Measurement Across Industries and Countries’, *American Economic Review*, 86 (5), 1216–38.

Böhm B. (1996), ‘Dynamic Econometric Specification and the Analysis of Structural Change’, *Dynamis*, Quaderno 2/96, IDSE, Milano.

Böhm, B. and L. F. Punzo (1994): ‘Dynamics of Industrial Sectors and Structural Change in the Austrian and Italian Economies, 1970–1989’, in Böhm B. and L. F. Punzo (eds), *Economic Performance. A Look at Austria and Italy*, Physica Verlag, Heidelberg.

Böhm, B. and L. F. Punzo (1995), ‘Structural Change in the Context of Uneven Regional Development. The Path of Italian Dualistic Economy Revisited with a New Dynamical Approach’, paper presented at the XI International Conference on Input Output Techniques, New Delhi, 1995.

Brida, G., M. Puchet and L. F. Punzo (2000), ‘Coding economic dynamics to represent regime dynamics’, WP no. 307, Dept. of Political Economy, Siena.

Day, R. H. (1992) ‘Complex Economic Dynamics: Obvious in History, Generic in Theory, Elusive in Data’, *Journal of Applied Economics*, 7, S9-S23.

Day, R. H. (1993), 'Non-linear Dynamics and Evolutionary Economics', in Day, R.H., and P. Chen (eds) (1993), *Non-linear Dynamics and Evolutionary Economics*, Oxford University Press, New York.

Day, R. H. (1994), (2001), *Complex Economic Dynamics*, vols. 1 and 2, The MIT Press, Cambridge, Mass.

Day, R. H. and Ping Chen (eds) (1993), *Non-linear Dynamics and Evolutionary Economics*, Oxford University Press, Oxford.

Doornik J. A. and D. F. Hendry (1994), PcFiml 8.0, Int. Thomson Publishing.

Durlauf, S. N. and D. Quah (1998) 'The New Empirics of Economic Growth', in Taylor, J. and M. Woodford (eds.), *Handbook of Macroeconomics*, Elsevier Science, Amsterdam.

Durlauf, S. N. and P. A. Johnson (1995), 'Multiple Regimes and Cross-country Growth Behaviour', *Journal of Applied Econometrics*, 10(4) 365-84.

Goldfeld, S. M. and R. E. Quandt (1972), *Non-linear Methods in Econometrics*, North Holland, Amsterdam.

Goodwin, R. M. (1947), 'Dynamical Coupling with Special Reference to Markets Having Production Lags', *Econometrica*; reprinted in Goodwin, (1982).

Goodwin, R. M. (1967), 'A Growth Cycle', in C.H. Feinstein (ed.), *Socialism, Capitalism and Economic Growth*; reprinted in Goodwin (1982).

Goodwin, R. M. (1982), *Essays in Non-linear Economic Dynamics*, Macmillan, London.

Goodwin, R. M. and L. F. Punzo (1987), *The Dynamics of a Capitalist Economy. A Multi-Sectoral Approach*, Polity Press and Westview Press, Cambridge and Boulder, CO.

Goodwin, R.M. (1990), *Chaotic Economic Dynamics*, Clarendon Press, Oxford.

Haavelmo, T. (1989) 'Nobel Memorial Lecture', in K. G. Mäler (ed.), *Nobel Lectures in Economic Sciences*, World Scientific Publishing Company.

Hansen B. E. (1992), 'Testing for parameter instability in linear models', *Journal of Policy Modeling*, 14, 517–533.

Hendry D. F. (1987), 'Econometric Methodology: A personal perspective', in Bewley T.F. (ed.) *Advances in Econometrics*, ch.10. Cambridge University Press, Cambridge.

Hendry D. F., A. J. Neale and F. Srba (1988), 'Econometric Analysis of Small Linear Systems using PC-FIML', *Journal of Econometrics*, 38, 203–226.

Hendry D. F., A. R. Pagan and J. D. Sargan (1983), 'Dynamic Specification', in Griliches, Z. and M. D. Intriligator (eds) *Handbook of Econometrics*, North Holland, Amsterdam.

Hicks, J. (1973) *Capital and Time*, Oxford University Press, Oxford.

Johansen, S. (1988), 'Statistical analysis of cointegration vectors', *Journal of Economic Dynamics and Control*, 12, 231–254.

Kaldor, N. (1957) 'A Model of Economic Growth', *The Economic Journal*, LXVII, December, 691-724.

Kaldor, N. and J. Mirrlees (1961), 'A New Model of Economic Growth', *Review of Economic Studies*, 29, 174–90.

Komiya, R., M. Okuno and K. Suzumura (eds) (1988), *Industrial Policy of Japan*, The University of Tokyo Press.

Lavezzi, A. (2000), 'Structural instability and unemployment: Evidence from the Italian regions', WP no.292, Dept. of Political Economy, Siena.

MacKinnon, J. G. (1991), 'Critical Values for Cointegration Tests', ch.13 in Engle, R. and C. W. J. Granger (eds), *Long-Run Economic Relationships*, Oxford University Press.

Maddala, D. S. (1977), *Econometrics*, McGraw Hill, New York.

Moriguchi, C. (1991) 'The Japanese Economy and Economic Structural Adjustments' *Economic Studies Quarterly* 42(1) (March), 1–11.

Moriguchi, C. (1995) 'Japan's Macro Economic Policy and Japan–US Economic Relations During the 1980s'. In Dutta, M. (ed.) *Economics, Econometrics and the Link*, Elsevier Science, Amsterdam.

Punzo, L. F. (1995), 'Some Complex Dynamics for a Multi-sectoral Economy', *Revue Economique*, 46, 1541–60.

Punzo, L. F. (1997), 'Structural cycles and convergence during processes of economic integration', published in Spanish as 'Cyclos Estructurales y Convergencia durante los procesos de integracion economica', *Revista de Economia*, Banco Central del Uruguay, 4(2) 3–5.

Quah, D. (1993), 'Empirical cross section dynamics in economic growth', *European Economic Review*, 37 (April), 426–34.

Ramsey, J. B. (1974), 'Classical model selection through specification error tests', in P. Zarembka (ed.), *Frontiers in Econometrics*, Academic Press, New York.

Scott, M. F. (1989), *A New View of Economic Growth*, Clarendon Press, Oxford.

Teräsvirta, T. (1996), 'Modelling Economic Relationships with Smooth Transition Regressions', Working Paper No.131, Stockholm School of Economics, November 1996.

Tong, H. (1990), *Non-linear Time Series. A Dynamical Systems Approach*, Oxford University Press, Oxford.

Yoshikawa, H. (1995), *Macroeconomics and the Japanese Economy*, Clarendon Press, Oxford.

## Appendix 1

### *The data base*

Data for Germany, France, and the USA are taken from the OECD publication, *National Accounts*, Vol II (various issues).

Value added by kind of activity (table 12), gross fixed capital formation by kind of activity of owner (table 3), employment by kind of activity (table 15).

For Germany all value added and investment data at constant prices have been converted into 1991 prices, for USA they have been converted into 1985 prices, and for France into 1980 prices.

Data for Italy have been constructed from two sources: for the period 1970 to 1980 they are from the ISTAT publication *Annuario di Contabilità Nazionale*, Tomo 2, tav1.7 etc. and converted into constant prices 1980. For data from 1980 until 1992 they are from the ISTAT data bank as published. Tests for the period 1980 to 1983 for both data sets have shown that the calculated growth rates are not affected by the systemic change in Italian national accounts.

The data for Japan come from regular publications by EPA, Economic Planning Agency.

*Table 3.A1* Coding of sectors

| *Sector* | *OECD (Germany, France, USA)* | *ISTAT (Italy)* | *EPA (Japan)* |
|---|---|---|---|
| Agriculture, hunting, forestry and fishing | 1 | 01 | 02 |
| Mining and quarrying | 2 | – | 03 |
| Manufacturing | 3 | | IND |
| Food, beverages and tobacco | 31 | 36 | 04 |
| Textiles, wearing apparel and leather industries | 32 | 42 | TEX |
| Wood, and wood products, including furniture | 33 | | |
| Paper, and paper products, printing and publishing | 34 | 47 | 06 |
| Chemicals and chemical petroleum, coal, rubber and plastic products | 35 | 17 | 07 |
| Non-metallic mineral products except products of petroleum and coal | 36 | 15 | 09 |
| Basic metal industries | 37 | 13 | 10 |
| Fabricated metal products, machinery and equipment | 38 | 24, 28 | MACH |
| Metal products, except machinery and transport equipment | | | 11 |
| Office and data processing machines, precision and optical instruments | | | PREC |
| Electrical goods | | | ELTR |
| Transport equipment | | 28 | AUT |
| Other manufacturing industries | 39 | 50 | 16 |
| Electricity, gas and water | 4 | 06 | 08,18 |
| Construction | 5 | 53 | 17 |
| Wholesale and retail trade, restaurants and hotels | 6 | 58 | 19 |
| Transport, storage and communication | 7 | 60 | 22 |
| Finance, insurance, real estate and business services | 8 | 69 | 20 |
| Community, social and personal services | 9 | 74 | 21 |
| Producers of Government services | 9 | 86 | 23 |

# Appendix 2

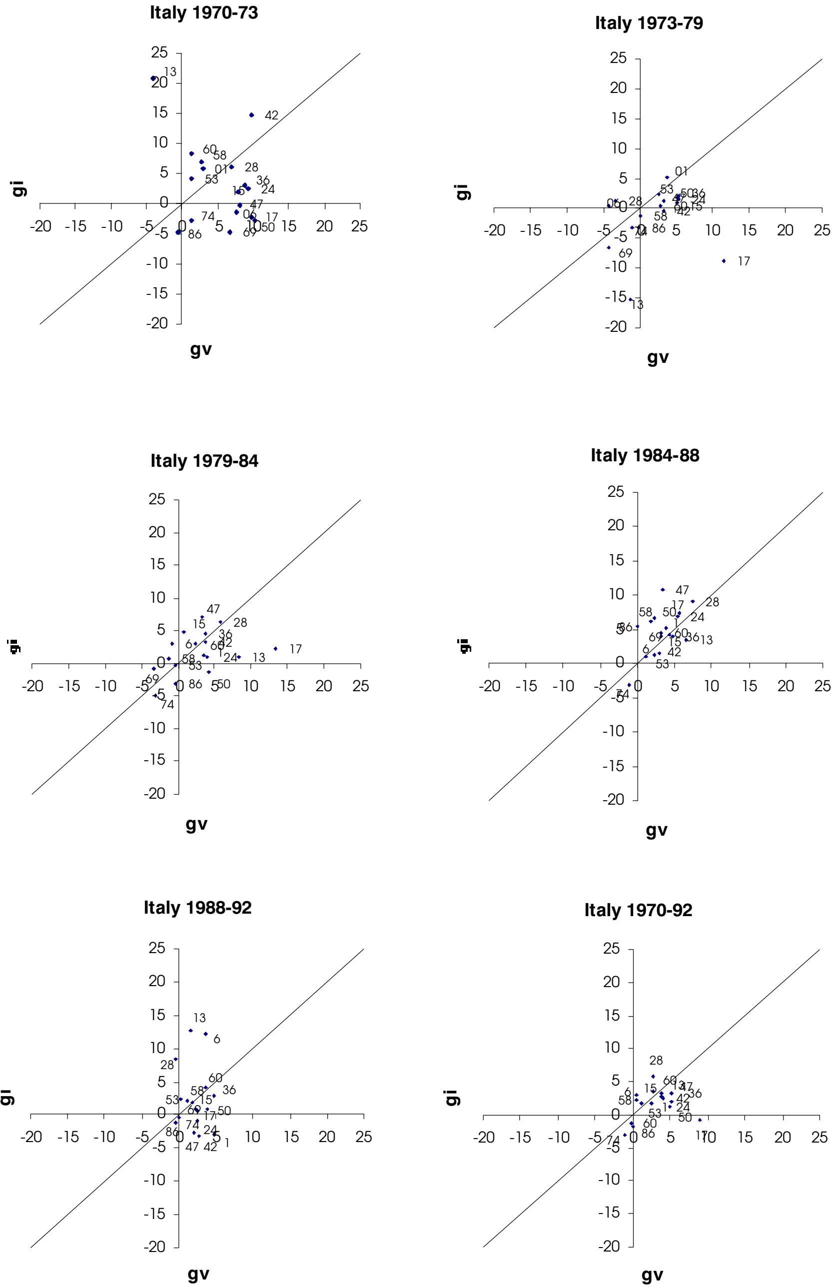
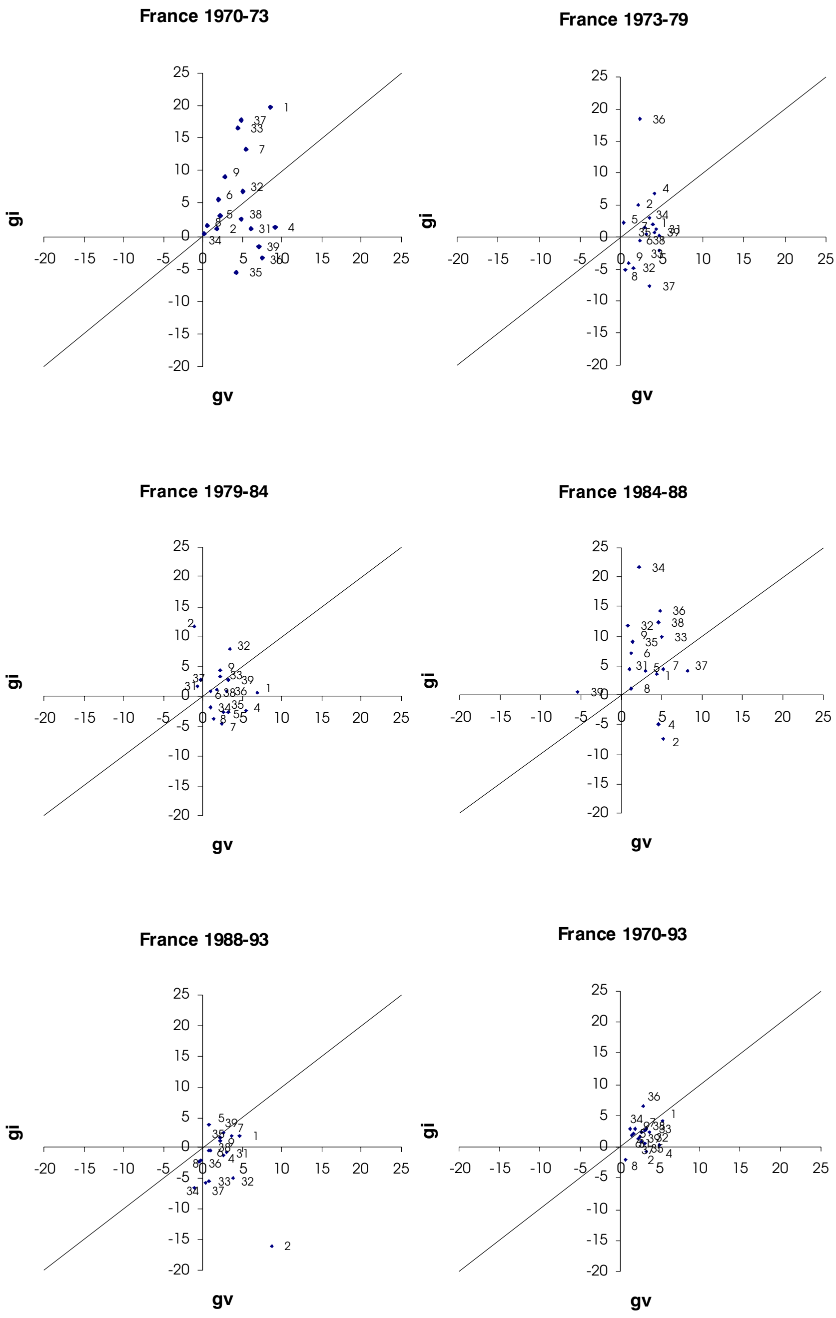


*Figure 3.A2.1* Italy



France 1970-73

France 1973-79

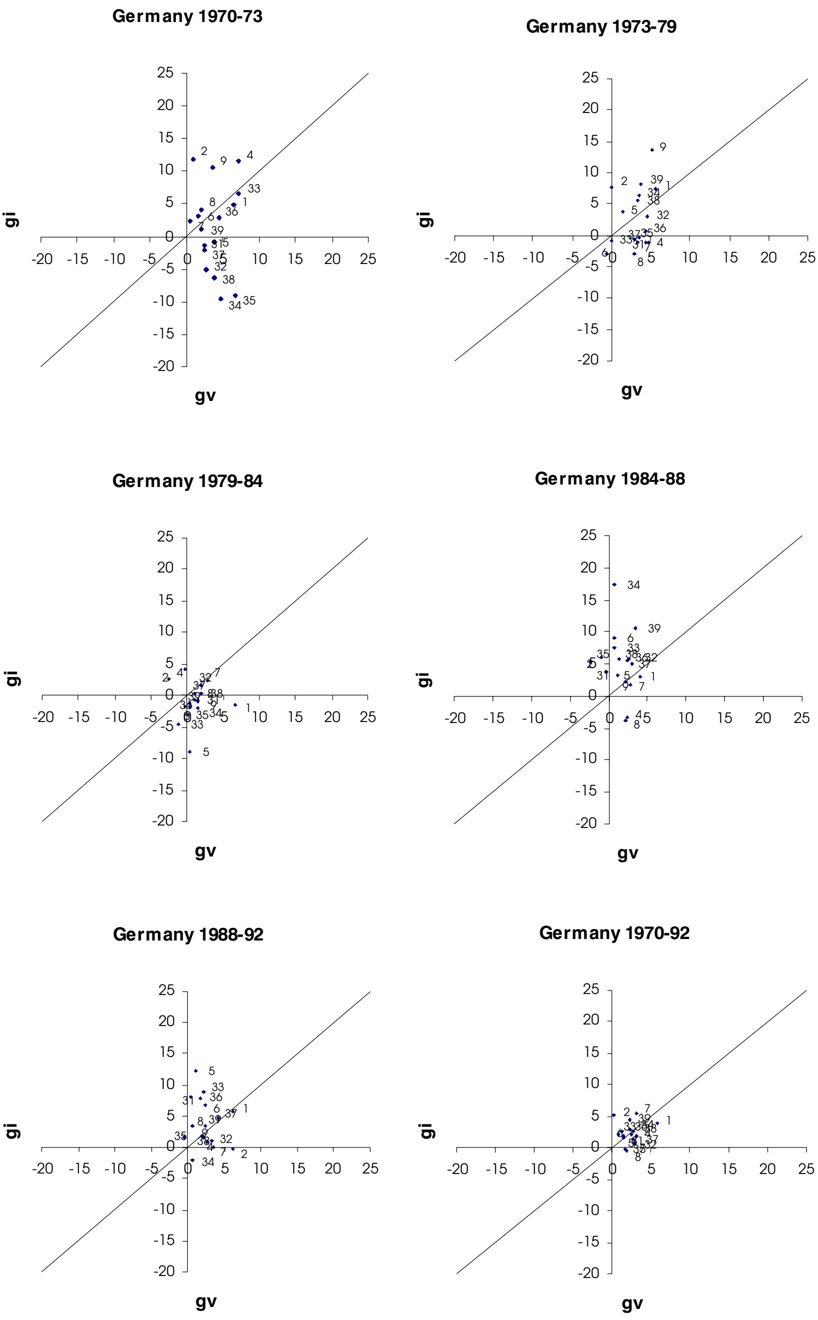
France 1979-84

France 1984-88

France 1988-93

France 1970-93

*Figure 3.A2.2* France



Germany 1970-73

Germany 1973-79

Germany 1979-84

Germany 1984-88

Germany 1988-92

Germany 1970-92

*Figure 3.A2.3* Germany

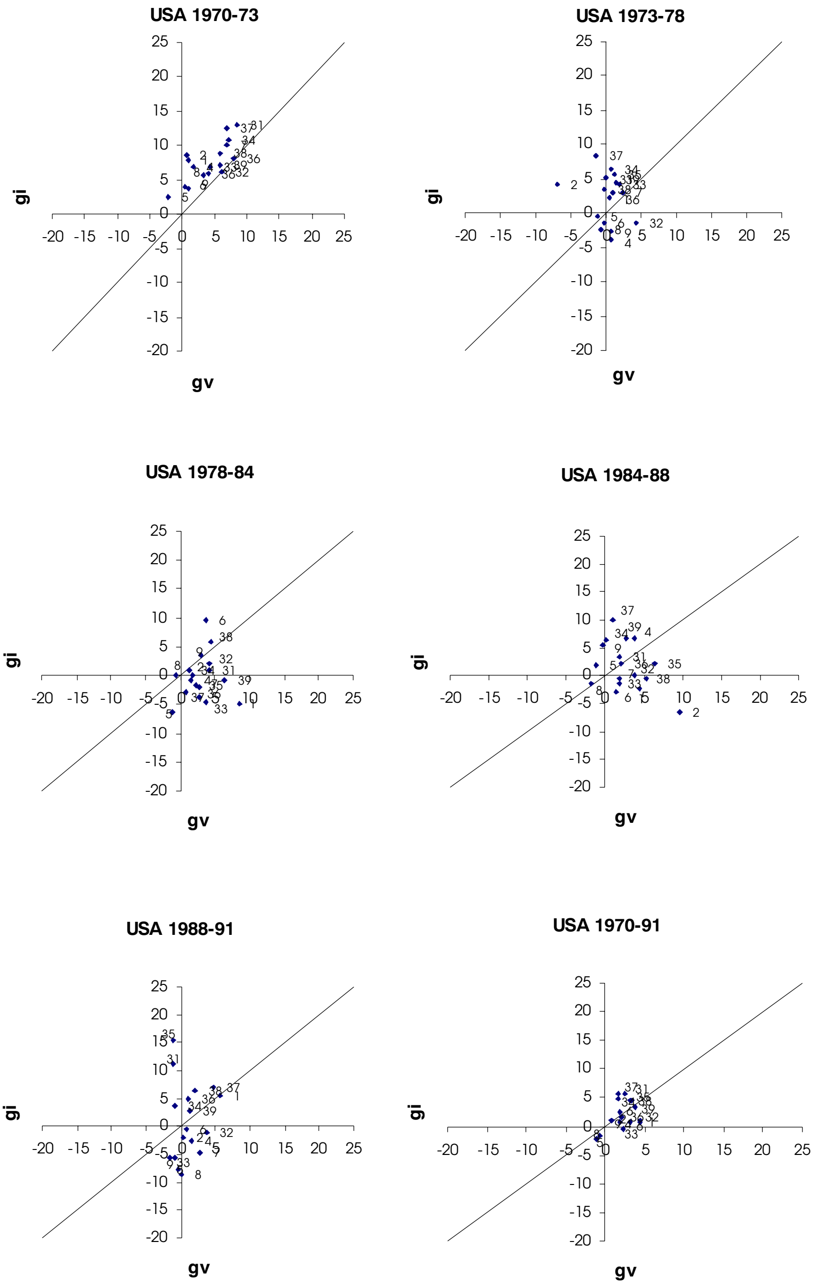


*Figure 3.A2.4* United States

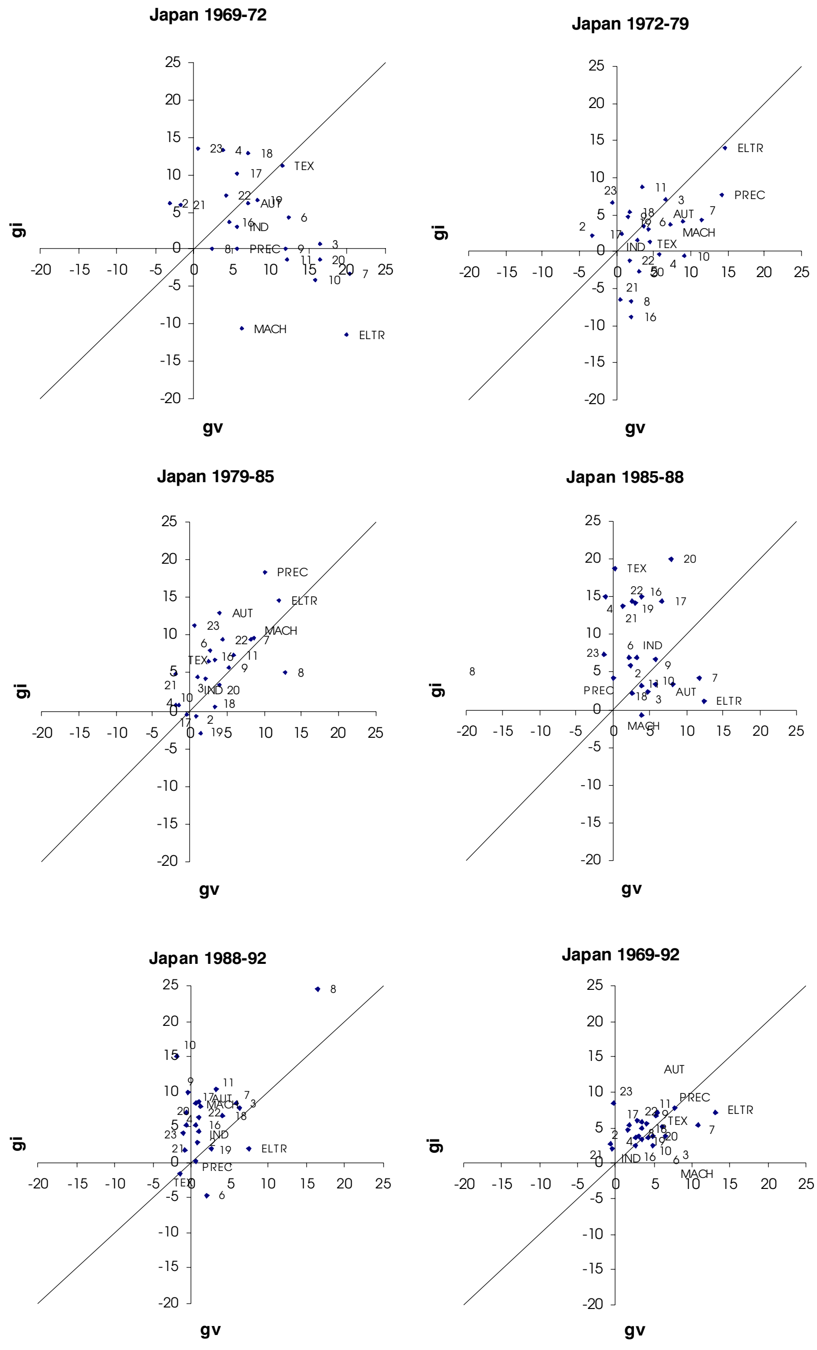


*Figure 3.A2.5* Japan

# Part II

# The macroeconomy and its dynamics

# 4 Qualitative dynamics and macroeconomic evolution in the very long run

*Richard H. Day and Oleg V. Pavlov*

When it is impossible to incorporate all the complex components involved, and it is therefore hopeless to represent behavior in accurate quantitative terms, it may still be possible to derive a compelling qualitative explanation of behavior. This possibility is illustrated in this chapter with the problem of explaining macroeconomic evolution. The resulting nonlinear, multiple–phase model exhibits trajectories that are qualitatively like those in the historical and archaeological records. Implications for understanding current transitions in the global economy and for development policy are suggested.

Dynamic macroeconomics is concerned with explaining growth, fluctuation, and structural change at aggregative levels. In this discussion our interest is with periods of time long enough to characterize fundamental changes in technology and economic organization that arise endogenously through the interaction of population, welfare, learning, and the internal and external economies and diseconomies of production.

The evolutionary character of the economic process is most easily seen at the microeconomic level when one kind of production technology displaces another. For example, exhibits in the old town museum of Århus describe in graphic detail the history of passenger transport in Denmark. Stage coach travel, which was introduced early in the nineteenth Century, was estimated to have produced some 8,300 passenger trips in 1833; 65,000 in 1846; and by 1859, 122,000. In the meantime, around mid-century the railroad had been introduced which, by the century's end, had all but eliminated the stage coach and was accounting for 15,000,000 passenger trips a year! Underlying this spectacular growth was a profound structural change in the nature of conveyance and human activity. Examples of similar growth waves abound throughout the industrial era.

Growth theory, as usually constituted, does not mimic the endogenous instabilities that drive an economy away from where it has been operating. In what follows a model of evolving macroeconomic structure is described that addresses these two aspects of very long run economic development, that is, (i) qualitative differences in economic structure and (ii) inherent instabilities.[1]

There is already a small literature on economic development in the very long run. Papers by Lee (1988), Kremer (1993), and Jones (1999) focus on the greater-than-exponential increase in human numbers, emphasizing demographic and productivity aspects of the story, but do not take into account *structural* changes in the process. Day and Walter (1989) introduce the role played by fundamentally different production technologies and social organizations based on discrete infrastructural requirements, the diseconomies that emerge as population grows within a given system, and the economies that can be achieved either by replicating economic units or by switching to a more ‘advanced’ system. Their approach emphasizes endogenous instabilities that drive the switch from one system to another. In a subsequent discussion of the model, David (1994) suggested incorporating structural changes induced by shocks that perturb a trajectory exogenously from a given basin into an alternative basin of attraction.

This study is a further development of the Day and Walter model that incorporates continuous learning by doing within given technological regimes. The model has been re-calibrated for European conditions with an emphasis on obtaining a better characterization of the long run qualitative trajectory of population and the structural episodes that underlie it.

Our approach requires a particular way of thinking about the facts which we call Qualitative Dynamics. It takes up as a general methodological issue in Section 4.2. In Section 4.3 the model components are described in nontechnical terms; Section 4.4 pulls all the parts together and the theoretical properties are summarized in Section 4.5. Section 4.6 is concerned with fitting the qualitative record from the hunting and food collecting stage of development to the present global information economy. A simulation example is presented that mimics features of development that stand out in the historical and archaeological records. Section 4.7 summarizes the contributions the qualitative econometric, multiple phase dynamics seems to offer. In the reference cited above, David rightly emphasize the fact that mathematical models such as this can only encompass a few of the factors necessary to explain the details of economic history. What they can do is capture in a stylized way some of the most obvious and important aspects of the historical process. In our opinion the present exercise amply demonstrates this potential.

## 4.1 Facts and qualitative econometrics

Economic science involves two distinctly different categories of fact: (i) facts about causality or how things work and (ii) facts about how states of the world change over time. The first set of facts provides the basis for a theory. A theory is then judged – from the scientific viewpoint – on how well it explains and predicts facts of the second kind. To put it slightly differently: (i) How well does a theory reflect facts of the first kind? (ii) How well does it ‘explain’ and ‘predict’ facts of the second kind?

When the focus is on microeconomic data, one can answer these questions in terms of facts garnered through direct observation. How and why things are done the way they are in particular industries or farming regions can form the basis for identifying decision criteria and causal structure. The resulting models can then be compared to data series at very low levels of aggregation obtained from individual firm records. Such was the case in the recursive programming models developed to simulate multi-product output, investment in alternative technologies and the changing patterns of resource utilization and productivity in various economic sectors.[2]

Micro detail is virtually impossible when the focus is on an entire national or world economy. Index numbers, aggregate variables or averages must be used to represent state variables. These are not observational variables but statistical artifacts that merely *reflect* economic activity as a whole. Indeed, if the concern is with periods of time long enough to reveal episodic structural changes, adequate numerical data do not exist at all; the quantitative record can – at best – only be pieced together from scattered fragments of information. It is here that recourse to *qualitative* as opposed to *quantitative* explanation is necessary: growth – not specific growth rates; fluctuation – not specific periodicity or statistical spectrum; stylized regimes – not specific technologies or institutions. Arranged in a time line, these qualitative facts become specific scenarios of development whose theoretical explanation can be contemplated.

Likewise, the specific attributes of the various technologies and institutions must be subsumed within coarse and abstract characterizations. We end up, then, not fitting a model to data or testing specific numerical hypotheses. Rather, we calibrate our model so that it can generate the qualitative development scenarios of interest. If this can be done on the *basis of components that reflect what we know about crucial economic structure*, then we have accomplished the task of *qualitative econometrics*. The work described here is undertaken in accordance with this methodological stance.

The qualitative facts of economic history writ large, so to speak, and that demand explanation include: (i) long run growth in population and production; (ii) intergenerational fluctuations in labor and welfare; (iii) secular increases in per capita productivity; (iv) periods of stagnation or declining productivity; (v) episodic regime switching involving fundamental differences in both socioeconomic organization and production technology; (vi) the replication and merging of similar economies, and (vii) the integration of economies to achieve a more advanced stage and the disintegration of an economy into a number of smaller economies at a less advanced stage of development. These various qualitative changes have occurred throughout history, as shall be summarized later.

## 4.2 Modeling macroeconomic evolution

### *4.2.1 The family function*

Time is reckoned in the classical unit of a human generation or a quarter century, each period is represented by a population of adults and their children who inherit the adult world in the next generation. Each generation must provide its own capital goods which only last the period. In what follows we measure population in terms of the number of individuals.

The number of children who survive to adulthood, $b$, depends on the average standard of living, $y$, on preferences (which may be presumed to depend on the broader aspects of culture) and on the survival characteristics of the social–technical regime. This relationship is called the *family function*,

$$b = g(y) := \begin{cases} 0 & , \quad 0 \le y \le \eta \\ (\alpha/q)[y-\eta] & , \quad \eta \le y < \zeta \\ 1+n & , \quad \zeta \le y \end{cases} \tag{4.1}$$

where $n$ is the maximal increase in population, $\eta$ is the birth/welfare threshold, and $\zeta$ is the threshold above which no further children are desired, or can survive, where $\zeta = \eta + (1+n)q/\alpha$.

The threshold standard of living below which no children survive to adulthood, exists for physiological reasons alone. Its empirical relevance is evident whenever famine or social conflict become severe, as has been the case recently, for example, in Africa and North Korea. A family function with such a threshold can also be derived from a reasonable specification of family preferences and constraints.[3]

### 4.2.2 Economies

To capture the character of economic development in the *very* long run, a number of factors, some of which have not been incorporated into macroeconomic growth models so far, have to be taken into account. We take these up now.

The organizational structure of a modern society includes households, producers, marketing firms, financial enterprises, and public institutions of various kinds. It provides the coherent framework of rules and procedures within which work can occur. It must be supported by human effort. The humans devoted to this effort form the infrastructure for a given socioeconomic system that mediates the human energy devoted to coordinating production and exchange, to providing social cohesion for effective cooperation, for training and enculturating the work force, and for producing the public goods such as waste disposal and public safety required for the well being of the work force. The knowledge that makes this human infrastructure effective is the *administrative technology*, a term due to Ester Boserup (1996). It must augment the production technology. As there are many institutions that are involved in the various infrastructural functions, the broader term *social technology* might be preferred.[4]

The adult, $x$, population is divided between the *labor force* engaged directly in production, $L$, and the *administrative* or *social workforce* that manages the infrastructure, $M$. In a decentralized economy much of this will be part of the private sector. A large part will also be part of the public sector. Both are necessary for a productive labor force. Given this, the number of adults in the infrastructure outside the family, $M$, and the number of adults in the labor force, $L = x - M$.

Diseconomies are, of course, implied by resource scarcity. They also accrue because of the increasing complexity of planning, communicating, and coordinating production as the economy grows. The ability to overcome them depends on the administrative technology and on the *social space* which this technology 'produces'. The social space defines the maximum number of individuals, say $N$, compatible with an effective socioeconomic order and with the feasible operation of the society's production process. *Social slack* is the difference between the social space and the current number of people, $S = N - x$. If there is positive social slack, then more people can be accommodated within the economy. As social space is 'used up', cooperation becomes increasingly difficult, social conflict increases, and productivity declines. When $S \leq 0$, society cannot function. Only when the social slack is positive can the society function. These *internal diseconomies* can yield absolutely diminishing returns to population within an economy.

Assume that the technological production function satisfies the usual assumptions:

$$Y = BG(L,S),\ L \geq 0,\ S \geq 0, \tag{4.2}$$

it is a continuous, strictly concave, homogeneous function of social slack and labor satisfying

$$\begin{array}{lclcll} G(0,S) & = & G(L,0) & = & 0 \quad \text{for all} \quad L,S & (a) \\ \lim_{L\to 0} \frac{\partial G}{\partial L} & = & \lim_{S\to 0} \frac{\partial G}{\partial S} & = & \infty. & (b) \end{array} \tag{4.3}$$

These assumptions imply that both labor and social slack are necessary for positive production and that both labor and social slack contribute positive but declining marginal productivities. The parameter $B$ is, as usual, the *total factor productivity level*. Substituting for $L$ and $S$, the production function can be written

$$H(x) := \left\{ \begin{array}{ll} 0 & ,\ x \in \backslash(M,N) \\ G(x-M,N-x) & ,\ x \in (M,N). \end{array} \right. \tag{4.4}$$

In words, output depends on the technology level, labor effect and slack effect.[5]

A given *economy* is characterized by (i) its (representative) family function (4.1), and (ii) its aggregate production function (4.4).[6]

### 4.2.3 Cultures

The key characteristic of the above formulation is that an economy based on a given social technology is bounded by its social space. At some point in the expansion of human numbers within an economy, the population may reach a level at which a new economy with the same system but with a newly constituted infrastructure can be split off in such a way as to increase welfare, in this way overcoming the internal diseconomies of population size. In effect, *the social space is increased by increasing the number of similar economies*. Contrastingly, if productivity were to fall sharply enough as the potential limit is approached, separate economies could *merge* to form a smaller number of economies with the same type of system, in effect *economizing on infrastructure*.

Here we assume these possibilities can be represented by *fission* and *fusion*, the former being the splitting of a given economy into two; the latter being the fusion of two or more economies to form a single one using the same basic system as before.[7] In addition to the internal diseconomies implied by resource scarcity and social space, external diseconomies should be recognized that derive from the total population of all the economies together. These diseconomies are, for example, caused by the environment's diminishing waste absorbing capacity as population expands and by the increase of the cost of extracting and refining resources as stocks decline. We assume that the *environmental capacity* depends on the production and administrative technology of a given system and can be expressed in terms of a maximum population density. Diminishing absolute returns to the work force can eventually come to pass as the total population becomes large.

For a given culture, the internal diseconomies of population can be overcome by replication; the external ones cannot. The aggregate effects of these diseconomies on production can be represented by a continuous *environmental damage function*

$$d = D(\bar{x} - x) \tag{4.5}$$

where $D(\bar{x}) = 1,\ D(0) = 0$ and $D'(\bar{x} - x) \leq 0$. The damage function reduces productivity as environmental capacity becomes progressively exhausted. We refer to $\bar{x}$ as the *environmental capacity* and the term $\bar{x} - x$ as the *environmental slack*. Once the world is full in the sense that *external* diseconomies become important, the replication of economies with the same basic structures must eventually come to an end.[8]

Out of all the conceivable numbers of economies that could exist by means of the fission/fusion process for a given system, we choose the one that is locally efficient for the population of a given size. We call a production function that optimizes the number of economies in the system the *cultural production function*. Given the environmental damage function, the locally efficient number of economies is

$$j = J(x) := \arg\max_{l \in \mathrm{N}^+} 2^l BH(x/2^l)D(\bar{x} - x). \tag{4.6}$$

Then the *environmentally constrained production function* is

$$K_{J(x)}(x) := 2^{J(x)} BH(x/2^{J(x)})D(\bar{x} - x). \tag{4.7}$$

If, when the environmental space is approached, a collapse occurs due to a very powerful drop in productivity, the population may reorganize itself by fusing groups into a smaller number of economies with the same technology as before. Then the stage is set for a new growth process through internal growth of the individual economies and through a resumption of replication in their numbers. Fluctuations in the numbers of economies, as well as in total population, could ensue, perhaps in a highly irregular way for a very long time.

Putting all this together leads to the concept of a *culture*: a society divided into a collection of similar economies, each based on the same socioeconomic system and whose total population is constrained by a common damage function. A small number of such economies can expand internally, then bifurcate, forming a new set of economies. This growth through fission can continue until the environmental slack is so diminished that productivity begins to fall in all the economies. Various possibilities follow, including population fluctuations and, if the productivity decline is sharp enough, *fluctuations in the number of economies* as well when existing groups merge to form a smaller number of economies, thus economizing on infrastructure and by this means re–establishing the possibility for resuming growth.

### *4.2.4 Cultural evolution*

Now consider the existence of several cultures, each characterized by the ingredients just described but each with distinct parameters. These alternatives represent different 'family values', 'ways of life', 'development blocks', distinct administrative and production technologies, and different environmental damage functions. We assume that the technologies have a natural order: each successive system in the order requires a greater overhead of human capital, possesses enhanced social and environmental spaces, and higher attainable production and population levels than its predecessors. We can therefore identify each distinct system by an index, $i \in \mathcal{T} := \{1,\ldots,\tau\}$.

Each culture is defined by the functions and parameters involved in its structure and by the set of potential alternative numbers of economies that can be replicated within its environmental constraints. A collection of potential cultures is a *cultural menu*.

In the face of limited environmental space for a given culture, a process of growth through replication of a given system cannot continue indefinitely. However, the *integration* and reorganization of infrastructure of existing economies could permit a jump to a more advanced system with a more demanding infrastructure if by doing so the environmental space of the new system is greater. Such a change in regime occurs in the present theory if average productivity is enhanced by doing so. This does not mean that each successive technology is uniformly more productive than its predecessor, but only that *at a given current total population* the switch to a new regime will enhance total factor productivity and, hence, the standard of living *at that population level*.

If in the process of growth, productivity falls enough as the regime's environmental capacity is exhausted and a higher regime is uneconomic or unavailable, existing economies could be forced to *disintegrate* into a larger number of smaller economies that require less elaborate infrastructures, in effect economizing on human capital by reverting to less infrastructure intensive regimes in a process of cultural reversion. *Uniform progress through the natural order of cultures cannot be presumed.*

Local efficiency must now be determined with respect to *both the culture and the number of economies adopting it*. Define

$$(i,j) = IJ(x,B) := \arg\max_{p\in\mathcal{T}} \max_{l\in\mathrm{N}^{++}} 2^l B^p H_p(x/2^l) D_p(\bar{x}^p - x). \tag{4.8}$$

where $B = (B^1,\ldots,B^\tau)$. Then the locally efficient culture is given by the culture, $i$, with $2^j$ similar economies. The production function that is locally efficient with respect to the selection of a culture and the number of economies using it, is given by

$$\begin{aligned} Y = K(x,B) \quad &:= \quad K_{IJ(x,B)}(x,B) := \\ &= \quad 2^l B^i H_i(x/2^l) D_i(\bar{x}^i - x),\ x \in X^{ij} \end{aligned} \tag{4.9}$$

where

$$X^{ij} = \{x \mid (i,j) = IJ(x,B)\}.$$

The set $X^{ij}$ is the set of population sizes for which $K_{ij}(x,B)$ is the cultural production function with the number of economies yielding the highest total (and average) output.

### *4.2.5 Learning by doing*

Output per unit of labor is assumed to increase as experience within a given regime accumulates, but continuing productivity advance based on experience or learning by doing can only occur within an 'active' system, that is, the one with the currently adopted basic technology (for only then can practical knowledge based on experience accumulate).

If $i$ is the system index identifying the active regime, then

$$\begin{aligned} B^i_{t+1} &= (1+\rho^i)B^i_t - \rho^i (B^i_t)^2/\tilde{B}^i && (a) \\ B^j_{t+1} &= B^j_t,\ j \neq i. && (b) \end{aligned} \tag{4.10}$$

Recall $B$ to be the vector of system technology levels, $B := (B^1, \dots, B^\tau)$. Given definition of the locally efficient culture (4.9), an operator,

$$T : (x_t, B_t) \longrightarrow B_{t+1}, \tag{4.11}$$

is defined by the $\tau$ equations (4.10). The $i^{\text{th}}$ component of the vector $B_{t+1}$ is generated by equation (4.10$a$) and the remaining coefficients by equation (4.10$b$). The process is asymptotically monotonically stable if $0 < \rho^i < 1$. We can think of $\tilde{B}^i$ as the *technology potential* for a given system.

The value of the technology parameter when a given regime is entered for the first time will be called the *innovating technology level*, denoted by the parameter $B_0$. Assuming that $0 < B_0 < \tilde{B}$, productivity will grow. The larger the potential and the smaller the innovating level, the larger is the technology gap; and the larger this gap, the more rapid the initial rate of productivity enhancement. As the stock of practical knowledge accumulates, the rate of accumulation eventually declines, and the technology level approaches its potential asymptotically. This process does not expand the ultimate limits on population allowed by the associated social and environmental spaces. That can be done in this model only by switching to a more advanced system.

## 4.3 GEM: A ‘general’ evolutionary model

### 4.3.1 The complete model

Putting all the above together, a ‘general’ evolutionary model, which we call ‘GEM’, emerges in which a society is portrayed as evolving through a sequence of alternative numbers of economies of a given type and switching among alternative socioeconomic systems in response to the standard of living of the current generation and to the potential standard of living that can be ‘selected’ through ‘self–reorganization’.

Let us summarize the model as a whole. The state variables of the system consist of the population, $x_t$, and the vector of productivity levels, $B_t = (B_t^1, \dots, B_t^\tau)$ with one element for each member in the cultural menu. The locally efficient culture and number of economies is given by

$$(i,j)_t = IJ(x_t, B_t), \tag{4.12}$$

defined by equation (4.8). Output is given by the cultural production function, defined by equation (4.9),

$$Y_t = K(x_t, B_t). \tag{4.13}$$

Average per capita standard of living is given by

$$y_t := \omega(x_t, B_t) = Y_t / x_t \tag{4.14}$$

and the succeeding population of adults by

$$x_{t+1} = g\Big[\omega(x_t, B_t)\Big] \cdot x_t \tag{4.15}$$

where $g(y_t)$ is defined in equation (4.1). With (4.11) we have, also,

$$B_{t+1} = T\Big[x_t, B_t\Big]. \tag{4.16}$$

Given initial conditions $x_0 = x,\ B_0 = (B_0^1, \dots, B_0^\tau)$, the $\tau + 1$ equations of (4.15)–(4.16) generate the trajectories for the state variables $(x_t, B_t)$.

### 4.3.2 The multiple phase characterization of trajectories

Defining the structurally dependent average standard of living, $\omega_{IJ(x,B)} := K_{IJ(x,B)}(x,B)/x$, we obtain the *phase zones*

$$\begin{array}{lll} \mathcal{D}^{0(i,j)_t} & := & \{x \in X^{ij} \mid 0 < \omega_{ij(x,B)} \le \eta^i\} \\ \mathcal{D}^{s(i,j)_t} & := & \{x \in X^{ij} \mid \eta^i \le \omega_{ij(x,B)} \le \zeta^i\} \\ \mathcal{D}^{n(i,j)_t} & := & \{x \in X^{ij} \mid \omega_{ij(x,B)} \ge \zeta^i\} \end{array} \tag{4.17}$$

Then at each time, $t + 1$,

$$x_{t+1} = \begin{cases} 0 & , \; x_t \in \mathcal{D}^{0(i,j)_t} \\ (\alpha/q)\left[2^{i_t} B_t^{i_t}(x_t - M^{i_t})^{\beta_{i_t}}(N^{i_t} - x_t)^{1-\beta_{i_t}} - \eta x_t\right] & , \; x_t \in \mathcal{D}^{s(i,j)_t} \\ (1+n)x_t & , \; x_t \in \mathcal{D}^{n(i,j)_t}. \end{cases} \tag{4.18}$$

Whenever $x_t$ enters a different phase zone, the equations governing $(x_t, B_t)$ change. This happens if the culture does not change $(i_{t+1} = i_t)$ but the number of economies increases or decreases due to fission or fusion does change, i.e. $j_{t+1} \neq j_t$. Or, it happens if a different culture is adopted so that $i_{t+1} \neq i_t$. In this case, the number of economies making up the new system will also change through integration or disintegration, as the case may be.

When $x_t$ enters a new phase zone, we denote the entry time '$s$'. Then $s_0 = 0$ and $s_k$ is the entry time for the $k^{\text{th}}$ regime switch. With this definition we can define an *episode* as the consecutive length of time periods governed by a given regime. With this definition $(i,j)_t$ is the same for $t \in \{s_k, s_k + 1, \ldots, s_{k+1}\}$. A sequence, $(i,j)_{s_k}, k = 0, 1, 2, 3, \ldots$ describes a trajectory in terms of the sequence of phases and phase zones through which it passes. Such a sequence is called a *scenario*.

Given this interpretation, the following additional endogenous variables are determined:

| | |
|---|---|
| aggregate production | $Y_t$ |
| average welfare per capita | $y_t$ |
| the time of entry into each episode | $s_k$ |
| the duration of each episode | $s_{k+1} - s_k$ |
| the governing regime in each episode | $(i,j)_t = IJ(x_t, B_t)$ |
| the governing number of economies using the dominant system | $2^{j_t}$ |
| the size of the aggregate infrastructure | $2^{j_t} M^{i_t}$ |
| the size of the labor force | $x_t - 2^{j_t} M^{i_t}$ |
| the social slack | $s_t = 2^{j_t} N^{i_t} - x_t$ |
| the environmental slack | $\bar{x}^{i_t} - x_t$ |

## 4.4 Theoretical properties

The mathematical analysis of the system is intriguing but requires a considerable amount of nomenclature to develop rigorously. A complete global analysis is given in Day (2000, Chapters 21–24). The upshot of it is that the conditions for various qualitative evolutionary scenarios to occur can be derived from an analysis of the structural properties of the model. To summarize:

(1) Evolution in this theory is driven by an unstable, deterministic process. (Of course, random shocks could be introduced but our purpose here is to isolate the intrinsic dynamics of the development process.)

(2) The probabilities of various possible historical scenarios can be derived in terms of sequences of qualitative events due to replication/merging and integration/disintegration.

(3) If the number of systems is finite and if the technology potentials of each system are not too high, then model histories must involve endless fluctuations, eventually sticking with a given regime or cycling in a nonperiodic fashion through an endless sequence of regimes with replication, merging, integration, and disintegration with jumping and reverting among systems.

(4) If there were a reachable regime with an asymptotically stable stationary state when its technological potential is reached, then the model's histories would very likely converge (with 'positive measure') possibly after many periods of local chaos to a classical equilibrium.

(5) Under some conditions trajectories can escape the zone of definition with positive measure. This corresponds to the demise of a society.

(6) If there is an upper bound on the environmental capacity for all systems, then the advancing regimes must become progressively 'squashed' against this ultimate bound.

(7) With large enough potential technology levels, the processes of output growth, replication, integration, and regime switching are accelerated as productivity is enhanced. Over the long run the system becomes more unstable. Development is less likely to get 'stuck' within a given regime. Economies are more likely to replicate or integrate and jump. Moreover, asymptotic convergence cannot then occur with positive measure. The chance of demise increases and the pressing of population against the environmental capacity occurs with ever greater speed, with an ever greater likelihood of collapse.

## 4.5 Fitting the evolutionary record

We now show that GEM can be calibrated so as to capture salient qualitative features of economic development listed at the end of Section 4.2.

### *4.5.1 The qualitative record*

Historians of the nineteenth century noticed that prior to the industrial take-off, economies had passed through distinct stages of development characterized by differences in production technology and in the organization of exchange and governance. Archaeologists, aided by modern methods of dating materials, began extending this picture backwards in time. By now they have constructed an approximate but coherent chronology of major developments on a worldwide basis that stretches back to the earliest evidence of a 'modern' human presence.

The great variety of human societies can be roughly grouped into a relatively small number of stages based on production technology and social infrastructure.

*Table 4.1* A time-line for the major systems

| *Index* | *Description* | *Permanent Entry: Generation* | *Duration: Generations* | *Duration: Years* |
|---|---|---|---|---|
| 1 | The hunting band | 1 | 3702 | 100,700–8,125 BC |
| 2 | Village agriculture | 3703 | 181 | 8,126–3,600 BC |
| 3 | The city-state | 3884 | 126 | 3,601–450 BC |
| 4 | Trading empires | 4010 | 89 | 451 BC–1775 AD |
| 5 | Industrial societies | 4099 | 8 | 1776–1975 AD |
| 6 | Global information economy | 4107 | 2 | 1976–Present |

To describe the major developments throughout the entire span of *Homo Sapiens Sapiens* and to take advantage of the known archaeological information, a reasonable minimal specification would be:

I. Hunting and gathering

II. Settled (village) agriculture

II. Complex societies and the city state (civilization)

IV. Trading empires

V. Industrial economies and the nation state

VI. Global information economies

In reality, various geographical areas traversed these stages at very different times and the advance through them did not increase uniformly from lower to higher index. Rather, progress from one to another, especially in earlier times, was interrupted by reversions to lower level stages. Moreover, fluctuations in income, population and capital have been typical. The overall picture is one of growth at fluctuating rates with sometimes smooth, sometimes turbulent transitions when jumps and reversions occurred until a 'higher' stage became firmly established. A summary of the archaeological and historical evidence concerning the transition through these stages and the various regime switching events is presented in Day (2000, Chapter 23).

A rough time line for the permanent transitions to the several stages is given in Table 4.1.

## 4.6 A simulation exercise

Given the qualitative patterns outlined above, we set for ourselves the task of calibrating the model so as to produce output, productivity and population trajectories through the six stylized systems according to the following scenario: population grows with intermittent fluctuations; numbers of economies within various systems increase (and perhaps decrease) through fission and fusion; economies integrate

*Table 4.2* System parameters for the mathematical history

| *Index i* | $\eta$ | $\alpha$ | $q$ | $n$ | $B$ | $\tilde{B}$ | $\rho$ | $\beta$ | $M$ | $SS(\equiv N)$ | $\delta$ | $\bar{x}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0.5 | 0.6 | 1 | 0.012 | 2.97 | 7 | 0.003 | 0.9 | 5 | 30 | 0.1 | 1.60E+06 |
| 2 | 0.5 | 0.6 | 0.95 | 0.015 | 2.055 | 10 | 0.02 | 0.6 | 250 | 2000 | 0.1 | 6.00E+06 |
| 3 | 0.45 | 1 | 1.5 | 0.022 | 2.2 | 12 | 0.12 | 0.6 | 2.50E+05 | 1.00E+06 | 0.1 | 1.90E+07 |
| 4 | 0.63 | 1 | 2 | 0.02 | 2.1 | 28 | 0.41 | 0.6 | 2.50E+06 | 1.00E+07 | 0.1 | 1.11E+08 |
| 5 | 0.7 | 1 | 2.2 | 0.19 | 2.1 | 60 | 0.4 | 0.6 | 2.00E+07 | 3.50E+08 | 0.1 | 4.00E+08 |
| 6 | 0.8 | 1 | 3 | 0.1 | 300 | 3000 | 0.45 | 0.4 | 3.98E+08 | 9.00E+08 | 0.1 | 9.00E+08 |

within a given culture to form smaller numbers of more advanced economies; economies disintegrate and revert of to a ‘lower’ stage. Parameter values, given in Table 4.2, were chosen by grid search so as to capture stylistic historical facts from Europe.

A simulation was begun with an initial population of 100 individuals $(x_0 = 100)$ and was continued for 4,168 periods or generations, a span of 102,700 years. Figure 4.1 shows the graph of population for this run – virtually a vertical line over the present, caused by explosive growth after a take-off a few centuries ago. In terms of sheer numbers, human population – relative to the present–appears utterly insignificant until the most recent centuries. This is the very long run trend explained in related ways by Lee, Kremer, and Jones in the studies referred to in Section 4.1. But the present model adds to this by identifying the action going on during the long duration when human numbers appear insignificant from the current perspective. This can be seen by transforming population numbers to logarithms as in Figure 4.2. In effect this change gives a heavy weight to small populations. It reveals some prominent features that are disguised within the thickness of the horizontal axis in Figure 4.1. Thus, at the beginning of the run population is seen to fall, then increase slowly for a long time. Next, it undergoes a considerable span of pronounced growth, reaches a plateau, then enters an era of very rapid

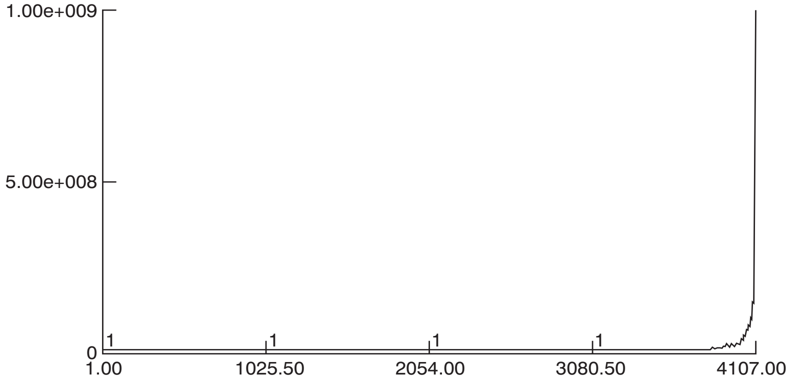


*Figure 4.1* A simulated population history (number of families). Time is measured in ‘generations’ of 25 years; population in number of individuals.

growth interspersed with fluctuations. Finally, the explosive growth of the recent past emerges.

When the data are plotted for shorter time spans, still more detail in the behavior of population emerges. This is shown in Figure 4.3. Panel (a) plots population for a span of some 3720 generations or 93,000 years. On this scale, irregular fluctuations of increasing magnitude appear. Panel (b) shows the next 388 generations or 9,700 years. There, too, periods of growth within the hunting and food collecting system are generated, interrupted by population decreases. Note the explosive trend of the most recent years.[9] Turning to the underlying economic forces, we find much more going on in terms of structural evolution. This can be seen in Figure 4.4, which displays the dominant system index, $i_t$, at each generation. These pictures reveal a model-generated history in which much of significance was happening economically, even though population was extremely small compared to the modern era.

The initial population in our simulation adopted the first system and remained with it for 3,702 generations. Growth occurred by means of the fission process, continuing for 10 millennia or so until the number of economies reached a level that persisted for many thousands of years thereafter. This long history of growth through replication underlies the population growth shown in Figure 4.3a.

The different time scale of Figure 4.5 reveals more detail. A system jump first occurs in period 3,630, some 11,975 years ago or about 9975 BC. It involves the integration of the very large number of hunting bands into a considerably smaller number of agricultural economies. They disintegrate within a generation, however, back into the original number of system 1 groups. Then structural fluctuation occurs, involving successive integrations and disintegrations, until the society locks into system 2. Growth then continues within this village agriculture system for 181 generations or nearly 4,525 years.

A similar round of structural fluctuations occurs between systems 2 and 3 (city-states) and systems 3 and 4 (trading empires), with corresponding fluctuations in the number of economies as the processes of integrations and disintegrations bring about system jumps and reversions. These outcomes reflect similar changes that are known to have occurred in reality, as briefly described above.

Once system 4 is locked in, reversions brought about by disintegration cease. After a few centuries, a jump to system 5 (the industrial revolution) occurs. The run terminates with a jump to system 6 (the global information economy), which in this simulation is interpreted as occurring in the last quarter of the twentieth century and takes place through an integration of industrial economies. The structural developments just described are shown in more detail by plotting the generated scenario for the most recent four millennia.

By the time the industrial system emerges, the number of autonomous economies has been drastically reduced. Nonetheless, the number of such states increases about the mid-nineteenth century and again about the mid-twentieth century. These are treated in the model as autonomous, but should be thought of not as isolated units but as interrelated yet identifiably distinct political entities.

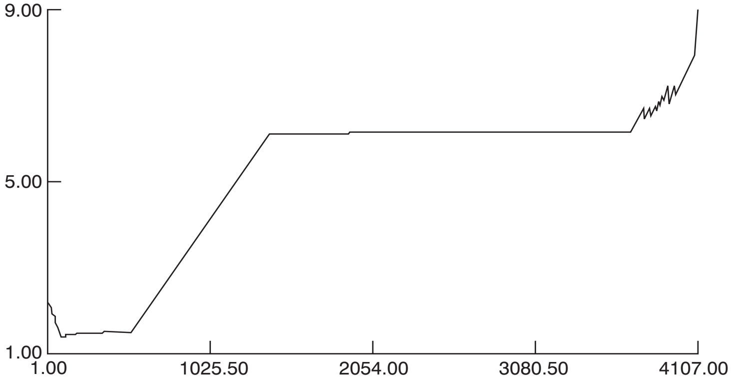


*Figure 4.2* Logarithm of simulated population

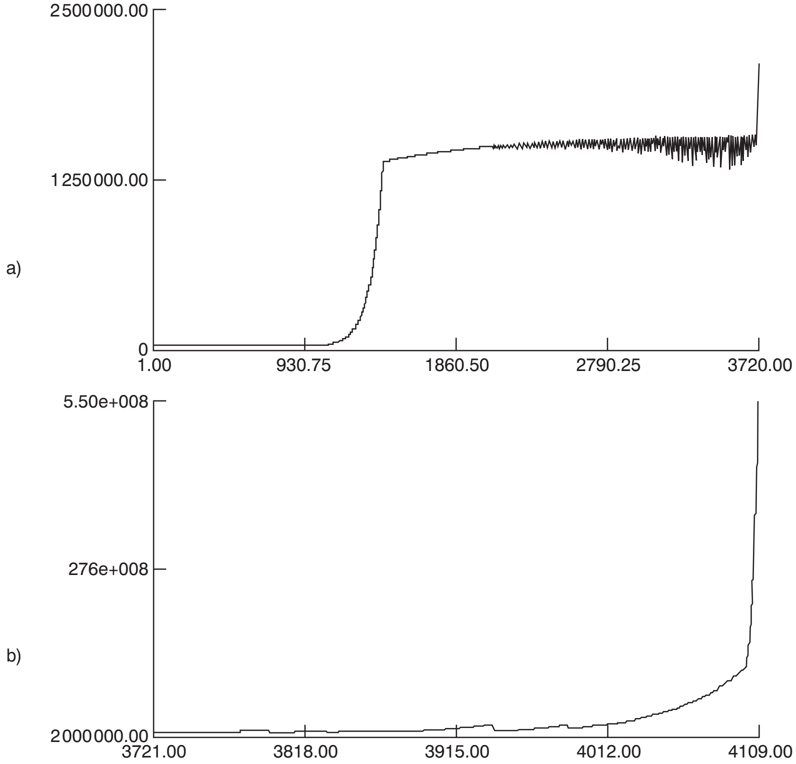


*Figure 4.3* Details of the population dynamics (number of families). Note the changing time scale from (a) to (b)

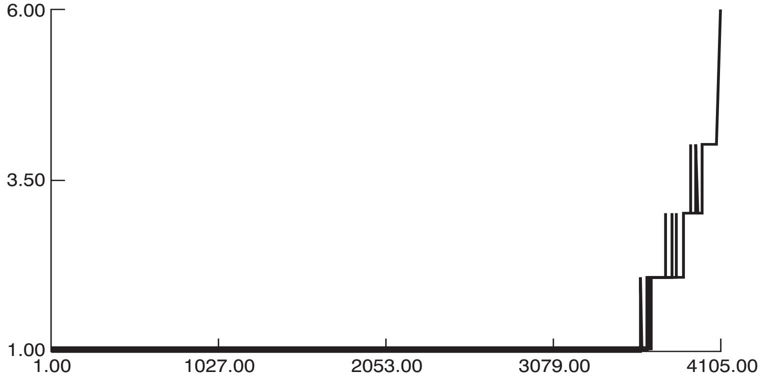


*Figure 4.4* Simulated history of structural change

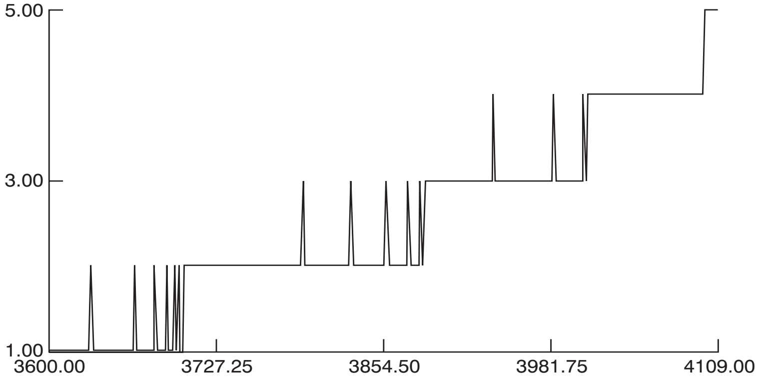


*Figure 4.5* Simulated history of structural change in greater detail

### *4.6.1 Conditional future scenarios*

The above simulation covers a span of history that brings the model society to the close of the twentieth century. Without adding a new breakthrough that would justify incorporating still another macro system but retaining exactly the same model structure as above, the simulation has been continued using as initial conditions the variable levels reached at the end of the first simulation, i.e. generation 4108 or year ‘2000’.

In this extended story, population divides into several ‘global systems’, which allow the population explosion shown to continue until the environmental space is so crowded that productivity is drastically lowered and the model economy experiences a disastrous fall in population. Using our modeling language, the society ‘self-destructs’ in some nine generations, which corresponds in real time to 2175 AD.

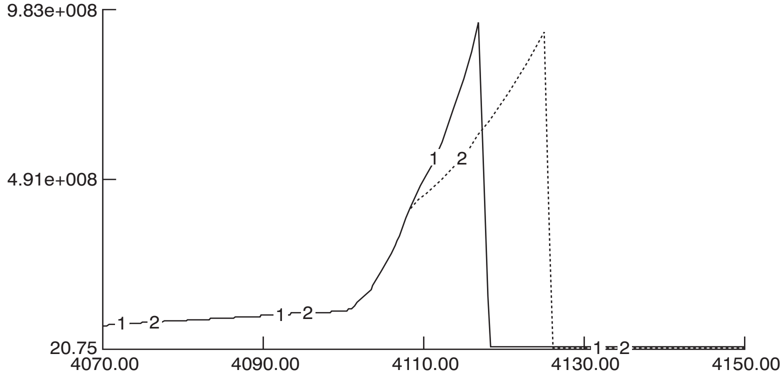


*Figure 4.6* Two extended simulations compared. 1 and 2 indicate number of families

*This extended simulation is not a forecast*. Rather, it is a conjecture about what *could* happen if no fundamental changes in technology or social parameters take place. Of course, we expect that such fundamental changes will take place.

In a second simulation all the parameters are retained except the natural growth rate '*n*' which is reduced to one half of the initial level. Without any advances in the technological culture, a mere reduction in the population growth extends the life of the model civilization to 17 generations after the first switch to the 'global economy'. The new 'end of the world' comes in 2400 AD. The situations with the two natural growth rates compared is shown in Figure 4.6. Obviously, attainment, of course, as long as the environmental capacity of the world is maintained, a 'zero population growth' rate, $n = 0$, would (in this model) extend the life of the civilization indefinitely.

## 4.7 Conclusion

Growth theory in the hands of Tinbergen, Solow, and Swan was designed to explain growth in the industrial countries during the first half of the twentieth century. Their models performed remarkably well and revealed the important roles of capital accumulation and productivity improvement. The elaboration of the theory by Lucas, Romer, and their followers incorporates the allocation of capital and human capital to productivity enhancement, in this way endogenizing the explanation of productivity advance in the long run, that is, over a century or more.

The approach described here emphasizes another fundamental aspect of economic growth that arises when the process is viewed over the *very* long run, that is, over millennia. That aspect is the discrete change in the production technology and social organization that has occurred along with the proliferation of economies with a given fundamental system, the fluctuation in their numbers, their unification in the transition to a system with a more elaborate social infrastructure, and their occasional disintegration into a larger number of economies with less elaborate infrastructures. The present theory characterizes this process and shows that parameter values can be chosen that fit the historical record in a qualitative sense, that is, that lead to model generated regime changes and fluctuations among the number of economies that are known to have occurred, according to a time line roughly in accord with the evidence.

This would seem to be the most that can be expected of such an exercise. Nonetheless, several important insights are suggested. First, population growth within a given fixed social system seems to have been limited, even given the presence of improving productivity. To overcome the limits, something more than doing the same thing better seems to be indicated. Overcoming the diseconomies that eventually emerge would seem to require basic changes in the way things are done, both in production and social organization. Endless growth, therefore, cannot be taken for granted, and collapses worse than a great depression can befall any culture that allows population growth to go unchecked while failing to reorganize itself to provide the means for maintaining coherence and symbiosis in its increasing numbers.

Our representation of cultural selection and phase switching does not explain the *process* by which such transitions are actually brought about. It only explains conditions sufficient to force such changes to occur. Both historian Quigley (1979) and archaeologist Flannery (1999) have discussed these processes in socio–political terms. Models like ours are very limited in what they can contribute to an understanding of the grand process of macroeconomic evolution. But they can provide a rigorous explanation of the demoeconomic forces that are involved.

## Acknowledgement

This chapter was presented as a paper at the Summer School on 'Economic Fluctuations and Structural Change' held in Siena, June 27–July 7, 1998. It draws on an earlier lecture presented at the Columbia University Conference, 'Managing Plant Earth', April 1997.

## Notes

1 Growth economists have, until recently, thought of the 'long run' in terms of centuries. Here we mean by 'very long run' growth over millennia.

2 See Day and Cigno (1978) and Day (1994). In the meantime, many alternative approaches have been and are being explored for studying microeconomic evolution.

3 The basis in household preferences for this Malthusian form is derived and the empirical basis for it reviewed in Day, Kim and Macunovich (1989). A more general version that incorporates a declining birth rate at high income levels is also introduced there. This feature, however, is not incorporated in the present model.

4 It must be emphasized that infrastructural functions are carried out in both public and private domains. The importance of the latter is sometimes overlooked. Large scale corporations allocate roughly half their expenditures on educational, research, managerial and administrative functions and roughly half on the production of goods and services. Although some economists would include such things in the category of intermediate goods used in the production process, it is worth distinguishing them

because their individual productivity cannot be measured in the usual ways (output per hour expended). Their productivity, like that of the elements of public infrastructure, is only reflected in the productivity of the entire organization. The contribution to the organization's success of individual scientists, teachers, managers, accountants is impossible to measure except by profit comparisons among similar organizations. Likewise, a productive public infrastructure will be reflected in some measure of aggregate accomplishment such as political, military or economic dominance, and/or a high level of culture and wide distribution of welfare.

Infrastructure has recently been receiving increasing attention. See North (1981) for very broad aspects and the World Bank (1994) for numerous details. For a suggestive attempt to quantify infrastructural effects on productivity, see the working paper by Charles I. Jenes and Robert Hall, 'Measuring the Effects of Infrastructure on Economic Growth', Stanford University.

5 The conditions expressed in (4.3b) imply that $\lim_{x \to M} H'(x) = \infty$, that $\lim_{x \to N} H'(x) = -\infty$, and that for all $x \in (M,N)$, $H''(x) < 0$, so $H(\cdot)$ is strictly concave on $[M,N]$.

6 Day and Min (1996) show that such an economy can display all the simple and complex possibilities: convergent growth, cycles, erratic fluctuations, and collapse. If continuous (exponential) productivity improvement is incorporated, then growth or fluctuations around a rising trend are possible or, as before, growth – possibly expanding fluctuations around a rising trend – followed by a collapse.

7 An alternative which allows for emigration and immigration is the process of 'shedding and assimilation' introduced in Day (2000).

8 For a discussion of the existence of an upperbound on population, see Cohen (1995a,b).

9 Necessarily, a different scale is used in each graph.

## References

Boserup, Ester (1996) 'Development Theory: An Analytical Framework and Selected Application', *Population and Development Review*, 22, 505–15.

Cohen, Joel (1995a) 'Population Growth and Earth's Human Carrying Capacity', *Science*, 269, 34–6.

Cohen, Joel (1995b) *How Many People Can the Earth Support?*, Norton and Co., London, UK.

David, Paul (1994) 'Do Economies Diverge? Comment on Day', in G. Silverberg and L. Loete (eds), *The Economics of Growth and Technological Change: Technologies, Nations, Agents*, Edward Elgar Publishers, Hants, UK, pp. 69–71.

Day, Richard H. and Alessandro Cigno (eds) (1978) *Modeling Economic Change*, Elsevier North–Holland, Amsterdam.

Day, Richard H., Kyoo–Hong Kim and Diane Macunovich (1989) 'Complex Demoeconomic Dynamics', *Journal of Population Economics*, 2, 139–59.

Day, Richard H. and Jean–Luc Walter (1989) 'Economic Growth in the Very Long Run: On the Multiple–Phase Interaction of Population, Technology, and Social Infrastructure', Chapter 11 in W. Barnett, J. Geweke, K. Shell (eds), *Economic Complexity: Chaos, Sunspots, Bubbles and Nonlinearity*, Cambridge University Press, Cambridge.

Day, Richard H. and Min Zhang (1996) 'Classical Economic Growth Theory: A Global Bifurcation Analysis', in T. Puu (ed.), *Chaos, Soilitons, and Fractals*, 7, 12, 1969–88.

Day, Richard H. (2000) *Complex Economic Dynamics, Volume II, An Introduction to Dynamic Macroeconomics*, The MIT Press, Cambridge, MA.

Easterlin, Richard A. (1978) 'The Economics and Sociology of Fertility: A Synthesis', Chapter 2 in Charles Tilly (ed.), *Historical Studies of Changing Fertility*, Princeton University Press, Princeton, NJ.

Flannery, Kent V. (1999) ‘Process and Agency in Early State Formation’, *Cambridge Archaeological Journal*, 9, 3–21.

Jenes, Charles I. and Robert Hall (1999) ‘Measuring the Effects of Infrastructure on Economic Growth’, Working Paper, Stanford University.

Jones, Charles I. (1999) ‘Was an Industrial Revolution Inevitable? Economic Growth Over the Very Long Run’, Mimeo, Department of Economics, Stanford University.

Kremer, Michael (1993) ‘Population Growth and Technological Change: One Million B.C. to 1990’, *Quarterly Journal of Economics*, 108,4, 681–716.

Lee, Ronald Demos (1988) ‘Induced Population Growth and Induced Technological Progress: Their Interactions in the Accelerating Stage’, *Mathematical Population Studies*, 1, 265–88.

North, Douglass (1981) *Structure and Change in Economic History*, W.W. Norton & Co., New York.

Quigley, Carroll (1979) *The Evolution of Civilization*, Liverty Press, Indianapolis, IN.

World Bank (1994) *World Development Report 1994, Infrastructure for Development*, Oxford University Press, Oxford.

# 5 Out-of-equilibrium dynamics

*Mario Amendola and Jean-Luc Gaffard*

## 5.1 Equilibrium dynamics

An instructive way to understand what we mean by out-of-equilibrium dynamics, and what kind of modeling it implies, is to compare it with the interpretation of economic theory as made explicit, e.g. in the real business cycle analysis. According to this view, the theory is reckoned to provide: 'an explicit set of instructions for building . . . a mechanical imitation system..to answer a question' (Kydland and Prescott 1996, p. 72).

The set of instructions consists mainly in behavioural axioms focusing on rational behaviour and concerning intertemporal optimization. Among them the axiom of technical efficiency, that is, to be always instantaneously on the optimal production function[1].

The system to be built by means of the above instructions is aimed at imitating a reality interpreted as an equilibrium state (on the assumption that the economy always really works as if it were in equilibrium), except for the case of stochastic shocks. The theory must then identify the stochastic properties of statistical series and reproduce them by means of a model. Modeling means describing scenarios which reproduce the 'reality' as defined above.

The shocks may affect production functions, thus bringing about exogenous cycles of a stochastic nature (Lucas 1980, Kydland and Prescott 1982), or they may affect the properties of production (and/or utility) functions, thus bringing about endogenous cycles of a deterministic nature (Grandmont 1985, Benhabib and Nishimura 1985). In both cases the 'fundamentals' of the economy are affected, and determine what happens to the economy itself.

## 5.2 Out-of-equilibrium dynamics

We do not consider axioms but formulate hypotheses such as to let dynamic economic phenomena stand out. In particular we believe that dynamics has to do with phenomena which are in the nature of *change*. Not 'quantitative' change, though – that is, a simple modification of the intensity of a given functioning of the economy – but 'qualitative' change, a change in the very *way* the economy actually operates.

Qualitative change is a structural phenomenon that can only take place through an adjustment. This implies a disruption of the productive structure on the operation of which the behaviour of the economy as a whole depends, and the shaping out of a different productive capacity. The adjustment thus comes down to an out-of-equilibrium process, characterized by the appearance of problems of co-ordination of economic activity which originate in the production side of the economy, since they actually reflect a breaking of the intertemporal complementarity of the production process as the result of the attempt to bring about a qualitative change (this will be shown more clearly in Section 5.4).

All this is canceled in the equilibrium analytical approach. Intertemporal optimization implies in fact intertemporal co-ordination. Equilibrium – as we have seen in the preceding section – is then no longer a position of the economy, but its very way of being (even when this takes the form of fluctuations which become themselves an expression of equilibrium). It becomes in a way the language through which economic theory expresses itself. We can thus no longer make the distinction between being in equilibrium or being out of equilibrium, a distinction which characterizes the analysis of the economists of the beginning of the century, like Wicksell or Marshall, and which still exists in Solow's model.

We look instead at equilibrium as at a state of the economy, defined with reference to given facts: namely, that the intertemporal complementarity of production, and hence the intertemporal co-ordination of economic activity, are assured. A breaking of this state implies the appearance of problems of complementarity and co-ordination, which throw the economy out of equilibrium

The co-ordination problems we are referring to are therefore different from the co-ordination problems of the standard approach, which are compatible with equilibrium. In this approach different co-ordination modes exist in relation to different informational contexts; co-ordination failures result then in the existence of multiple sub-optimal equilibria, not in the breaking of equilibrium.

This different perspective implies a different view of the very role of economic theory.

We no longer aim at *describing scenarios*, at reproducing 'realities' interpreted as the different facets of a given way of being of the economy. We focus on a particular state of the economy, its being out of equilibrium, which is in the nature of a process. Our aim is then *to interpret this process*, to analyze the salient moments and links which make it up.

## 5.3 The analytical implications of a change of focus

The above change of focus has momentous analytical implications. In the first place the usual distinction between a long term, where equilibrium obtains, and a disequilibrium short term disappears. A process is neither a short nor a long term: it is a sequence of disequilibria which link on and shape the evolution of the process itself.

The analysis of this process, on the other hand, does not call for a traditional type of model, that is, a model capable of generating a 'solution' in the sense of a type of behaviour of the economy (the attainment of a point or the following of a path) characterized by certain specific features (efficiency, optimality, and so forth). What we are after, instead, is to follow the evolution of the economy, traced out step by step by the above mentioned sequence of disequilibria, in order to investigate its viability. The essence of a thorough process, in fact – when the focus is on the process itself, as it should, not on its point of arrival which can't even be defined abstracting from how the process builds up along the way – is in its going on, that is, in its being viable. The concept of 'solution' interpreted as a given configuration of the economy, therefore, has no meaning when we refer to an out-of-equilibrium process, unless by this we mean the sorting out of the conditions which, step by step, make it viable. This calls for a monitoring of the process itself to bring to light its salient moments: which can only be achieved by means of numerical experiments, that is, by simulations that, under certain conditions (chosen so as to stress aspects relevant to the analysis) allow to unveil what happens 'along the way'.

In this light also the usual distinction between the terms 'exogenous' and 'endogenous' must be interpreted in a different way. In a model there are variables and there are parameters which reflect the existing constraints. In the standard analysis the constraints, which exist outside and above the economy and which determine its behaviour, are taken to be exogenous.

> But once we recognize that the time over which change takes place is a continuing and irreversible process which shapes the change itself, as we have to do when we consider a qualitative change, 'it is impossible to assume the constancy of anything *over time* . . . The only truly exogenous factor is *whatever exists at a given moment of time*, as a heritage of the past'. (Kaldor 1985, p.61) In the analysis of an out-of-equilibrium process . . . we thus have to consider as a parameter, and hence as exogenous, not some given element chosen beforehand by reason of its nature or characteristics, but whatever, at a given moment of time, is inherited from the past. What appears as a parameter at a given moment of time is therefore itself the result of processes which have taken place within the economy: processes during which everything – including resources and the environment, as well as technology – undergoes a transformation and hence is made endogenous to the change undergone by the economy. Thus, while the standard approach focuses on the right place to draw the line between what should be taken as exogenous and what should be considered instead as endogenous in economic modeling – a line that moves according to what we want to be explained by the model – out of equilibrium . . . the question is no longer that of drawing a line here or there but rather one of the time perspective adopted. Everything can be considered as given at a certain moment of time, while everything becomes endogenous over time.
>
> (Amendola and Gaffard 1998, pp.32–3)

Finally, it must be stressed that the 'fundamentals', which determine the equilibrium values of the relevant magnitudes of the economy, no longer play the same role out of equilibrium, when the focus is on a process rather then on a given configuration of the economy. Different evolution paths can be associated in fact to given fundamentals, according to how the out-of-equilibrium process actually evolves, and the fundamentals themselves undergo a change during this process, given the very definition of qualitative change. The 'fundamentals', in other words, are no longer fundamental.

## 5.4 Intertemporal complementarity and co-ordination

A viability problem arises, out of equilibrium, due to the appearance of co-ordination problems. There is no question of viability when co-ordination is assured, as it is the case in equilibrium. Co-ordination problems arise, during the adjustment process stimulated by the attempt to bring about a qualitative change, as the result of what happens in this case to productive capacity – namely, what happens sequentially *in time* to productive capacity.

Consider a shock which affects the regular behaviour that defines an equilibrium state of the economy. The immediate effect of this shock is to throw the economy itself out of equilibrium. This depends on a modification in the structure and the functioning of the underlying productive capacity which actually determines the behaviour of the economy. The focus therefore must be put in the first place on the production process.

This can be illustrated by considering a neo-Austrian representation of the process of production (Hicks 1973, Amendola and Gaffard 1988, 1998), that is, by portraying it as a fully vertically integrated process (where labour is the only primary input contemplated) taking place through a sequence of periods which make up a phase of construction and, following it, a phase of utilization of productive capacity. Although in equilibrium this representation comes down to a representation in terms of standard production functions, it also allows us to show what happens out of equilibrium, when the time dimension of production comes to the fore, with all its analytical implications. In neo-Austrian terms an equilibrium structure of productive capacity is represented by an array of production processes in the (different periods of the) phase of construction and of the phase of utilization which are consistent with each other, in the sense of supporting a steady-state of the economy. This age structure of productive capacity implies not only a horizontal dimension, the number and age structure of production processes at each given moment of time, but also a vertical dimension, the time pattern of production consistent with the former. When this is so, not only construction and utilization, but also the economic activities behind these phases, investment and consumption and supply and demand of final output, are consistent with each other, at each moment of time and over time. The complementarity over time of the production process (that is, the complementarity between construction and utilization) implies the co-ordination over time of the decision (and allocation) process. Production, as to its effects, is synchronized.

As we have mentioned, the attempt to bring about a qualitative change, that is, to modify the existing behaviour of the economy, throws the economy itself out of equilibrium in that it results in a breaking of the intertemporal complementarity of the ongoing processes of production. This, we have also seen, implies the appearance of problems of co-ordination, as saving, investment and consumption, as well as supply and demand of final output, go out of balance in correspondence to the fact that the phases of construction and utilization of productive capacity are no longer consistent with each other. The time dimension of production becomes relevant.

Reaction to these disequilibria, and the adjustments of productive capacity aimed at re-establishing the consistency over time of construction and utilization disturbed by the original shock, stimulate an out-of-equilibrium process that propagates the initial distortion over time without needing any further shock. What happens then to the economy must be looked at as a process sketched out step by step by sequentially interacting disequilibria which engender a complex dynamics. The backbone of this process is the accumulation through which adjustments, which necessarily imply a restructuring of productive capacity, take place in time.

A modeling of this process, besides the above mentioned neo-Austrian representation of the production process, requires the hypothesis of an adaptive behaviour of agents, so as to stress the sequential character of the decision process and its interaction over time with the production process. A sequence 'constraints–decisions–constraints', fed by complementarity and co-ordination problems interacting over time, sketches out the evolution path of the adjustment process undergone by the economy (see the model expounded in Section 5.6 which gives analytical structure to this argument).

## 5.5 Analysis and policy conclusions

Out-of-equilibrium dynamics, we maintain, consists in the analysis of the above defined adjustment processes to shocks implying structural modifications, that is, altogether different behaviours of the economy.

We shall now consider the processes associated with specific shocks, selected so as to show not only how the analysis itself is carried out but how it may lead to policy conclusions which are often just the opposite to those resulting from standard equilibrium analysis.

Under so called rational expectations, an economy originally in a steady state which faces a forward biased technological change (i.e. with reference to a neo-Austrian framework, a technological change implying higher construction costs more than compensated for by lower utilisation costs) converges towards a new steady state characterized by a higher level of productivity and higher real wages. The unemployment and the productivity slowdown which appear as a consequence of the emergence of a human resource constraint are transitory (Amendola and Gaffard 1998, pp. 157–8). As a matter of fact, the hypothesis of rational expectations makes it possible to maintain the consistency over time of construction and utilization and hence hampers the problems linked to the intertemporal complementary of production processes from appearing. With more realistic assumptions about firms' behaviours, the evolution of the economy is more complex and depends on the way co-ordination issues are dealt with. In what follows, on the one hand we shall assume that firms also have 'rational expectations' in the sense that they try to maintain an investment behaviour aimed at preventing the distortions of productive capacity from occurring. Investment thus will only be constrained by the availability of financial/or human resources. But on the other hand we shall assume that firms have an adaptive behaviour in the sense that they determine their current supply in each period in reaction to market disequilibria perceived in the previous periods.

Three different scenarios of evolution will be analysed hereafter, each one corresponding to a different monetary policy.The first one is characterized by a growth rate of money supply which is maintained equal to the current growth rate of the economy (a Friedman rule). The second one is characterized by a monetary policy aimed at keeping the price level stable (the rule that could be applied by an independent central banker). The third one is characterized by a discretionary monetary policy implying a temporary increase in the growth rate of money supply. In each case 900 simulations have been performed corresponding to different values of the prices and wages reaction coefficients randomly chosen in the interval [0–1.5].

In the first scenario (Figures 5.1), 99 per cent of viable paths end with lower real wages, lower productivity and persistent unemployment. In other words the co-ordination conditions prevent the economy from capturing the benefits of the new technology. Things get worse in the second scenario (Figures 5.2): 100 per cent of viable paths end with lower real wages, lower productivity and persistent unemployment; moreover the final level of these variables is lower than in the previous scenario. On the contrary, in the third scenario (Figures 5.3), 95 per cent of viable paths end with higher real wages and higher productivity, and 98 per cent end with full employment. It is worth mentioning that an increase in the saving rate (a decrease in the take out, as defined in Section 5.6) in scenarios of the first category does not improve the results (Figures 5.4): only 80 per cent of the simulated paths are viables; 83 per cent of the viable paths end with lower real wages, 91per cent with lower productivity, and 91 per cent with persistent unemployment.

These numerical experiments help to solve some puzzles in modern growth economics. First, they provide an original and robust explanation of the so called productivity paradox. As a matter of fact we have been able to show that the productivity slowdown which, at the economy's level, can result from the introduction of a new and superior technology must be attributed to the co-ordination failures arising during the out-of-equilibrium process of adjustment to the technology itself. These experiments also throw a different light on the role played by monetary policy in a growth process. In a context where this process is properly seen as an out-of-equilibrium adjustment whose main requirement is viability, monetary policy cannot be neutral. It must be aimed at keeping the evolution of the economy within a stability corridor which assures viability. Finally, these experiments confirm that the saving rate does not really matter. More precisely, an increase in the saving rate does not make sure that potential productivity gains will be realized.

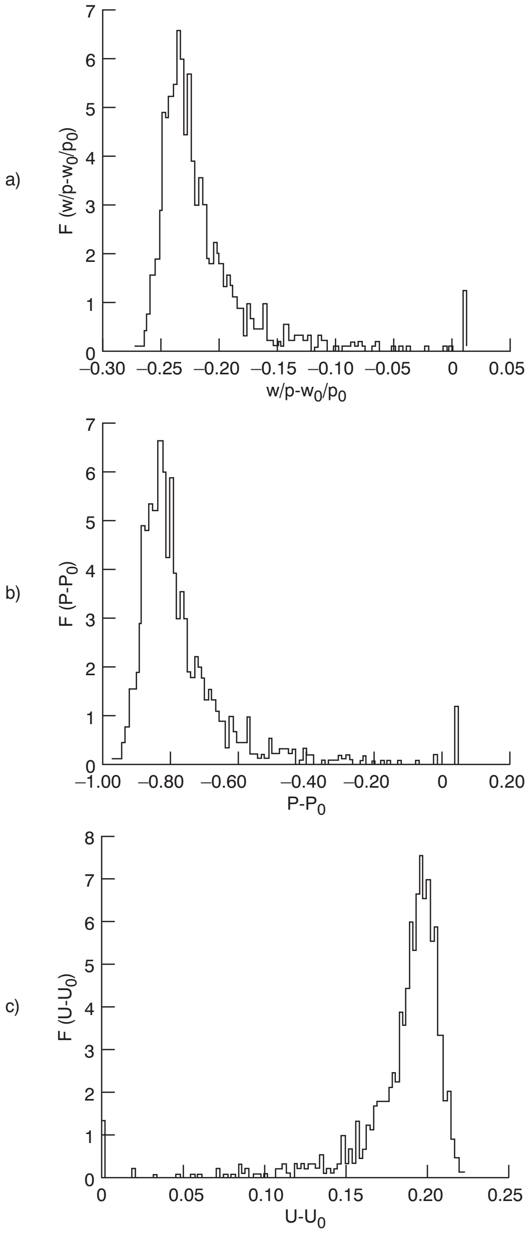


*Figure 5.1* Scenario 1: a) distribution of real wages, b) distribution of productivity, c) distribution of final employment.

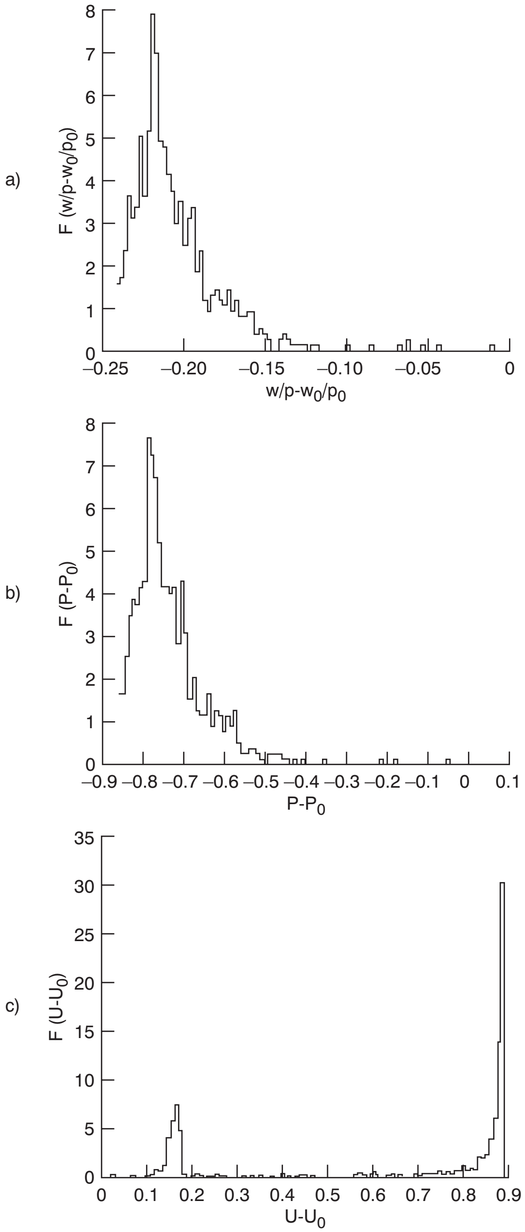


*Figure 5.2* Scenario 2: a) distribution of real wages, b) distribution of productivity, c) distribution of final employment.

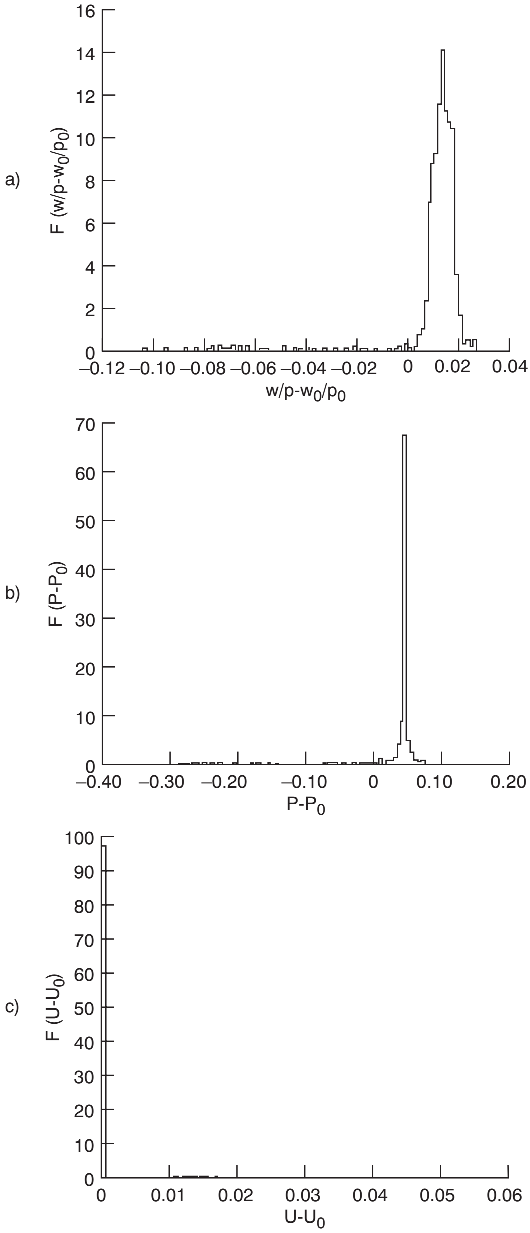


*Figure 5.3* Scenario 3: a) distribution of real wages, b) distribution of productivity, c) distribution of final employment.

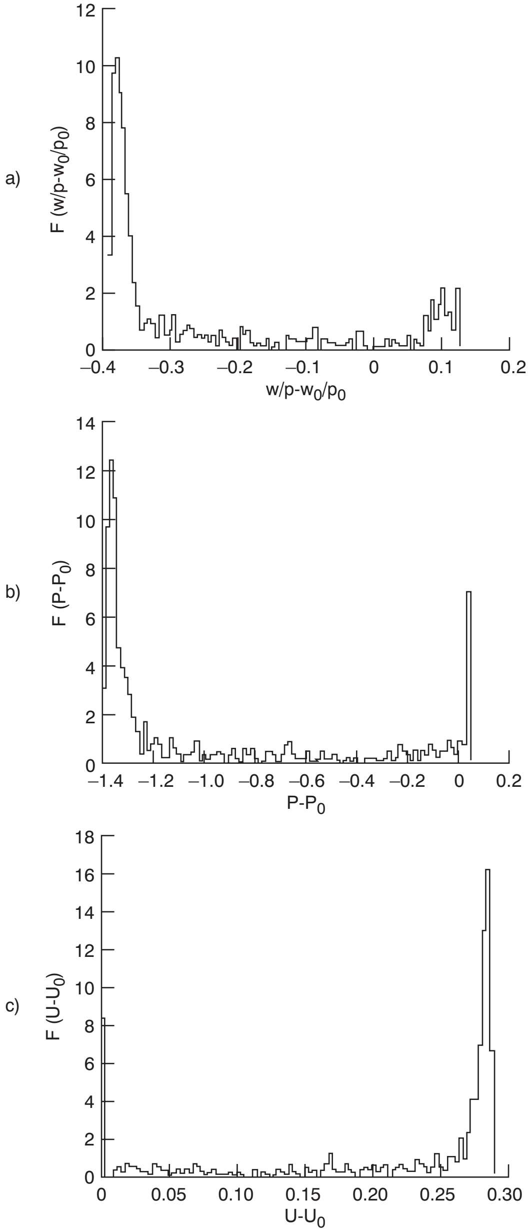


*Figure 5.4* Scenario 1 with increase in saving rate: a) distribution of real wages, b) distribution of productivity, c) distribution of final employment.

On the contrary, it brings about intertemporal complementarity problems that may result in cumulative imbalances and erratic fluctuations in the growth rate of the economy which hamper its viability.

## 5.6 Appendix

The following model which has been used for the simulation analysis whose results have been summed up in the figures in the text, is based on a neo-Austrian representation of the production process, where

$\mathbf{a} = [a_1, \ldots, a_{n^c}, a_{n^c+1}, \ldots, a_{n^c+n^u}]$

is the vector of labour input coefficients, and

$\mathbf{b} = [b_{n^c+1}, \ldots, b_{n^c+n^u}]$

is the vector of final output coefficients, in the different periods of the phase of construction $(1, \ldots, n^c)$ and of the phase of utilization $(n^c + 1, \ldots, n^c + n^u)$.

Productive capacity at time (t) can be written

$\mathbf{x}^c(t) = [x_1(t), x_2(t), \ldots, x_{n^c}(t)]$

$\mathbf{x}^u(t) = [x_{n^c+1}(t), x_{n^c+2}(t), \ldots, x_{n^c+n^u}(t)]$

where $\mathbf{x}^c(t)$ and are $\mathbf{x}^u(t)$ the vectors of the production processes in the phase of construction and the phase of utilisation, respectively.

In a steady state:

$x_j(t) = x_{j-1}(t-1) = x_{j+1}(t)G$

$x_j(t) = x_N(t)G^{N-j}; N = n^c + n^u; j = 1, \ldots, N$

where G=1+g is the growth factor.

The labour supply is equal to:

$L^S(t) = L^S(0)[1+g]^t$

where g is the natural growth rate.

The resources required to carry out production and to sustain consumption are financial resources. There are ‘external’and ‘internal’financial resources.

$m(t) = \min[p(t)s(t), p(t)d(t)]$are the money proceeds from sales (internal financial resources), where s and d are the supply and demand of final output and p its price. $f(t)$ is money supply (external financial resources)

In a neo-Austrian model, the wage fund represents the resources which sustain the process of capital accumulation. It is determined by the minimum between the available financial resources $F(t)$ and the wage fund constrained by the available human resources:

$\omega(t) = \min[F(t), w(t)L^s(t)]$

with

$F(t) = m(t-1) + h^f(t-1) + f(t) - c(t)$

where c(t) is the take out (the resources withheld from financing production processes) and $h^f(t)$ the monetary idle balances of finance which pile up when the human resource constraint is more stringent than the financial constraint:

$h^f(t) = F(t) - w(t)L^s(t)$

The decisions are taken as follows:

*Production decisions* The current production is the difference between the current supply and the stocks actually put back on the market. It cannot be greater than the existing output capacity inherited from the past.

$$q't) = s(t) - o(t-1) \leq \mathbf{b}\widetilde{\mathbf{x}}(\mathbf{t-1})$$

where $o(t) = s(t) - d(t)$. It determines the vector of production processes in the phase of utilisation $\mathbf{x}^{u}\mathbf{(t)}$

The money value of current supply is determined on the basis of expected money proceeds

$$p(t)s(t) = Em(t)$$

*Consumption decisions* The money value of current households' final demand is determined by their financial constraint

$$y(t) = p(t)d(t) = \omega(t) + c(t) + h^{h}(t-1)$$

where $h^{h}(t) = \max\left[p(t)\left(d(t) - s(t)\right), 0\right]$ are the monetary idle balances of households which pile up when the value of final demand exceeds the value of current supply.

*Investment decisions* All the available financial resources can be invested:

$$i(t) = \omega(t) - \omega^{u}(t)$$

where $\omega(t) = w(t)\mathbf{a}^{u}\mathbf{x}^{u}(\mathbf{t})$.

Alternatively the investment can be determined as follows:

$$i(t) = \min\left[\omega(t) - \omega^{u}(t), w(t)\mathbf{a}^{c}(\mathbf{t})\mathbf{x}^{c}(\mathbf{t})\right]$$

where $x_1(t) = x_{n^c}(t)\left[1 + g^{*}(t)\right]^{n^c+1}$ and g is the growth rate which makes it possible to prevent distortions in the age structure of productive capacity.

*Price and wage decisions* Price and wage change from one period to the next in reaction to market disequilibria

$$g_p(t) = \kappa\Phi(t-1)$$

$$g_w(t) = \nu\Psi(t-1)$$

where $g_p(t)$ is the rate of variations of price, $\Phi(t-1)$ the rate of excess demand for the final ouput, $g_w(t)$ the rate of variation of wage, $\Psi(t-1)$ the rate of excess demand for labour.

*Control variables* The money supply is alternatively determined on the basis of the current growth rate

$$f(t) = f(t-1)\left[1 + g_m(t-1\right]$$

or in such a way to keep the price level stable

$$f(t) = f(t-1)\left[1 + g_f(t-1) - \varsigma g_p(t)\right]; \varsigma > 0$$

or on the basis of a targetted growth rate

$$f(t) = f(t-1)\left[1 + g^{\star}(t-1\right]$$

The take out is alternatively determined on the basis of the current growth rate

$$c(t) = c(t-1)\left[1 + g_m(t-1\right]$$

or in such a way as to gradually increase the amount of resources devoted to finance investment

$$c(t) = c(t-1)\left[1 + g(t-1)\epsilon^{t-t_c}\right]$$

where $t_c$ is the date at which the rate of take out starts decreasing.

## Notes

1 It may be noted that this does not allow the typical problems connected with the phenomenon of production (its time dimension, the dissociation of inputs and output and hence of costs and proceeds over time, etc.) to stand out at the analytical level. This, we shall see in what follows, represents the real watershed between an equilibrium and an out-of-equilibrium analytical context.

## References

Amendola, M. and J-L. Gaffard (1988) *The Innovative Choice* Oxford: Basil Blackwell.

Amendola, M. and J-L. Gaffard (1998) *Out of Equilibrium*, Oxford: Clarendon Press.

Benhabib, J. and K. Nishimura (1985) ‘Competitive Equilibrium Cycles’, *Journal of Economic Theory*, 35, 284–306.

Grandmont, J.M. (1985) ‘On Endogenous Competitive Business Cycles’, *Econometrica*, 53, 995–1046.

Hicks, J.R. (1973) *Capital and Time*, Oxford, Clarendon Press.

Kaldor, N. (1985) *Economics without Equilibrium*, Cardiff University College Press.

Kydland, F.E. and E.C. Prescott (1982) ‘Time to Build and Aggregate Fluctuations’, *Econometrica*, 50, 1345–70.

Kydland, F.E. and E.C. Prescott (1996) ‘The Computational Experiment: an Econometric Tool’, *Journal of Economic Perspectives*, 10, 69–85.

Lucas, R.E. (1980) ‘Methods and Problems in Business Cycles Theory’, *Journal of Money, Credit and Banking*, 12, 696–715.

# 6 Disequilibrium growth in monetary economies

## Basic components and the KMG working model

*Peter Flaschel*

## 6.1 Introduction

This chapter surveys, also for the general reader interested in non-market clearing models of growth and fluctuations, the foundations – and the core K(eynes)M(etzler)G(oodwin) model built on them – of the general framework underlying joint past and present work with Carl Chiarella and others on integrated disequilibrium models of monetary growth. The core KMG model of disequilibrium growth and its analysis is founded in this chapter on specifically reformulated and extended partial dynamic models of the literature, the PC-AC approach (of Goodwin and Rose) and the IS–LM–PC approach (of the textbook literature). We also briefly indicate at the end how the fundamental KMG model thus obtained (with its six basic laws of motion) can be extended into the direction of fairly detailed, high-dimensional macrotheoretic disequilibrium growth models of monetary economies, with strong relationships to a variety of models currently used for structural macroeconometric model-buildings and their applications.

In order to indicate the scope and perspective of the macrodynamics to be considered in the following let us briefly discuss here the following graphical representation of the essential components of an integrated Keynesian disequilibrium growth theory. Figure 6.1 shows in the middle what might be considered the backbone of Keynes's (1936) General Theory, the basic causal nexus, that makes goods markets behavior (via the investment decisions of firms) dependent on what is achieved on financial markets, and labor markets in turn on the outcome on the goods markets with their determination of output through expected sales (and intended inventory changes), which in turn depend on actual aggregate goods demand (and actual inventory changes).

In the center of interest in the macroeconomic debate of the last two decades has been, however, quite a different module of the macroeconomy, representing so-called supply side features or, as we prefer to characterize it, the dynamic wage–price block of a fully-specified approach to macrodynamics. This block is shown as wage–price spiral bottom left in Figure 6.1 and it is surely no exaggeration to state that it has dominated the development of macrostatic and macrodynamic mainstream thinking in the last decades, even to the extent that it was claimed that

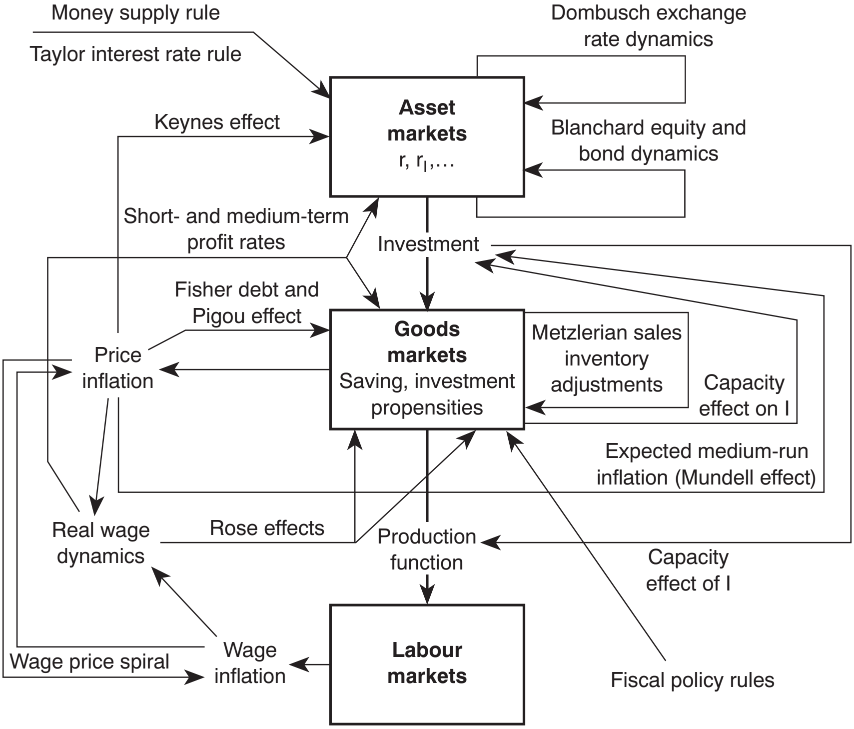


*Figure 6.1* The scope of traditional Keynesian theory

the Keynesian IS–LM model when integrated with supply side effects and rational expectations degenerates to a supply side dynamics without effective demand constraint, see Sargent (1987, Ch.5), and for a critique of such results Flaschel (1993) and Flaschel, Franke and Semmler (1997).

The consequences of such a concentration on supply side issues has been that neither the relevance of the above considered causal nexus, nor the feedback structure (or repercussions) we have added to it in Figure 6.1 (the arrows leading back into the middle area) have received much interest. It may indeed be claimed, that interest in the many stabilizing or destabilizing feedback chains known from the literature on Keynesian macrostatics or macrodynamics (of the 1960s and the 1970s in particular), was nearly absent in more recent mainstream macrodynamic model building with their wage–price interactions, up to the use of stabilizing Keynes- and Pigou-effects, which often simply served the purpose of providing for the deterministic part of the models a shock-absorber scenario on the basis of which impulse-propagation mechanisms could be erected and investigated. Even if the Keynesian demand constraints were given some attention, there was thus generally the scenario of rapid convergence to full employment steady states unless shocks continued to hit the economy.

Yet, there are (locally) destabilizing effects of rising inflation and expected inflation (Mundell-effects), there are destabilizing Metzlerian inventory accelerator mechanisms, there is the implication (when aggregate demand depends on income distribution and the real wage) that either price of wage flexibility must be destabilizing (normal or adverse Rose-effects), there are destabilizing Fisher debt deflation effects with respect to investment or consumption behavior (in particular if there is high debt of firms or certain types of households). There are cumulative processes in exchange rate and asset market dynamics (where expected and actual rates of appreciation or depreciation do exercise a positive feedback on each other). There are therefore a variety of reasons on the basis of which one might form the expectation that steady states of integrated Keynesian or disequilibrium growth models are more likely surrounded by centrifugal forces than by centripetal ones, implying the necessity to introduce additional nonlinearities should the dynamics depart too much from the steady state due to these forces.

The study of growth with endogenously generated fluctuations, based on the disequilibrium adjustment processes and resulting feedback chains of the just characterized type and coupled with additional nonlinearities they may give rise to far off the steady state, thus should remain on the agenda of macrodynamic theorizing, if only as a (not yet) well-established and well-known alternative scenario with which achievements of other integrated approaches to monetary growth can be compared. There is of course always the still fashionable possibility, to model economic dynamics such that the jump variable technique can be applied, which by definition removes from sight all instabilities initially present in the dynamics, so that there remains not much to be compared. Models employing the jump variable technique have however stressed the importance of a treatment of anticipated future events, a topic that should also be considered and solved by the integrated disequilibrium growth dynamics we have in mind.

In Figure 6.1 we finally present (top-left and bottom-right) the addition of policy feedback rules to the considered interaction of a structure of market dominance with the wage–price spiral and the shown feedback structures of the private sector, yet not necessarily with the understanding that policy can just manipulate this scenario from the outside, but that there may be interactions with the behavior assumed for the private sector and the type of policy considered to a more or less significant degree. Policy issues are however not yet well developed in the approaches we shall review in this chapter and thus only supplement here our summary of traditional Keynesian model building, or better what it should have been, but has not been yet, at least from the dynamically fully integrated point of view.

Our perspective in the following is thus integrated Keynesian disequilibrium growth analysis as it is obtained, on the one hand, from prototypic models of fluctuating growth and, on the other hand, of inflation and stagflation. This implies in a natural way the inclusion of four of the feedback structures just discussed, with particular stress on the dynamics of wages and prices, but also of quantities and thus on goods and labor market reaction patterns. Due to space limitations we arrive here however only at a fundamental or working model type of integrated disequilibrium growth theory, while further extensions (open economies, financial markets, policy rules and more) remain for future research.[1]

In Section 6.2 we will consider supply side dynamics in isolation, yet not one of the AS variety, but one that extends the basic approaches of Goodwin (1967) and Rose (1967), here called AC-PC analysis, towards more refined treatments of wage–price dynamics, and also towards an inclusion of adverse adjustments of real wages and real debt (in the case of deflation in particular). Section 6.3 will then consider the textbook IS–LM model augmented by a certain type of wage–price dynamics (based on Friedman's views on full employment and the role of inflationary expectations). We will show there that the resulting dynamical system is far from being well understood, and that it will give rise more likely to persistent fluctuations in employment and inflation rates rather than to the shock absorber behavior that is generally believed to be the outcome here. We also show that the labor market NAIRU may be a goods market NAIRU in fact.

Section 6.4 considers modern discussions of wage or price Phillips curves and tries to offer a unifying framework for such discussions (to be integrated into our systematic evolution of integrated models of disequilibrium growth later on). This section therefore demonstrates that much remains to done when AC-PC analysis is extended to include Keynesian goods market dynamics. We also show that a special case drawn from this general framework of wage–price dynamics may give rise to an endogenous explanation of the NAIRU which is quite different from the one that rules the roost in the literature on this rate of (un)employment.

In Section 6.5 we then present and investigate our basic modeling of integrated disequilibrium growth theory, the core or working KMG model of disequilibrium growth obtained as endpoint of a systematic discussion of models of monetary growth of orthodox type in Chiarella and Flaschel (2000b), namely the Tobin type models, the Keynes–Wicksell model types, the IS–LM growth model (without and with smooth factor substitution) and IS–LM growth models based on the (incomplete) dynamic multiplier story. This section briefly presents theoretical results on the resulting six-dimensional dynamics (non-existing so far in the literature on this level of generality). We here study Rose-, Mundell- and Metzler-effects in their interaction and thus obtain, on the one hand, an already fairly advanced feedback structure between goods and labor markets (including the conventional Keynes-effect as short-cut to the interaction with asset markets). In comparison to Figure 6.1, on the other hand, the working KMG model still provides only a limited picture of the working of actual economies (on the macrolevel). The reader is referred to Chiarella and Flaschel (2000h) for the consideration of extensions of this fundamental integrated disequilibrium growth model towards the other topics included in figure 6.1 and also towards a theoretical penetration of modern structural macroeconometric model building of the type shown in the figure.

## 6.2 AC-PC analysis

In static or dynamic models of the supply side of Keynesian and other types, the marginal productivity hypothesis, the equality of real wages with the marginal product of labor, or of the price level with marginal nominal wage costs, appears in many respects as the central element on which analyses of supply side processes are to be founded, by way of a conventional Aggregate Supply (AS-) curve as the competitive theory of the price level in a Keynesian setup (where firms are to be treated as quantity-takers and price-setters) or in a Walrasian setup as theory of labor demand, giving rise there to a Lucas type supply function with price-taking firms. Yet, at least in a Keynesian environment with price setting firms, supply side processes should have an explicit and detailed representation even in the case of fixed proportions in production, where the marginal productivity theory is no longer applicable, since this type of modeling of macroeconomic interactions should not depend on the assumption of so-called neoclassical production functions (though its implications may vary to some extent with the assumptions that are made with respect to available production technologies).

We shall show in this chapter that attempts to include the dynamics of supply in models with a Keynesian short-run give rise to an analysis of wage–price dynamics that can be usefully compared with the implications of the supply side dynamics of Goodwin's (1967) model of 'the growth cycle', and its extensions, which we will discuss in this section in simple graphical terms. In further sections we will exemplify that this claim indeed holds with respect to integrated Keynesian models of traditional type. In order to provide a lively idea of the dynamics we have in mind when reconsidering integrated models of disequilibrium growth later on, we will present the Goodwin model here in even simpler terms than were used in the original work of Goodwin, the fundamental and prototypic nature of which has been stressed by Solow (1990) in particular.

After our brief representation of Goodwin's (1967) seminal contribution we will indicate what has to be added to it when integrated into a Keynesian framework as far as goods and asset markets are concerned. We then provide two extensions of the Goodwin model (and their synthesis) which serve to indicate its potential for more advanced types of analyses. The first of these extensions is related to the goods-market analysis of Rose (1967), another early contribution to cyclical growth and employment fluctuations which at least in some respects is closely related to the Goodwin approach. The second integrates stocks and financial assets in a very fundamental way in order to add the topic of debt deflation to the price dynamics of the Rose approach.

### *6.2.1 AC-PC model building: the basic structure*

Figure 6.2 provides the basic elements needed to derive the Goodwin (1967) overshooting growth cycle mechanism in an environment where we are still abstract from technical change. We are implicitly assuming a fixed proportions technology. We have top left a real wage Phillips Curve (PC), relating the rate of growth $\hat{\omega}$ of real wages $\omega$ with the state of the labor market, expressed by the rate of employment $V$. This curve has been drawn as strictly convex, but it needs in the minimum only fulfill the following three conditions in order to obtain the conclusions of the Goodwin (1967) model:[2]

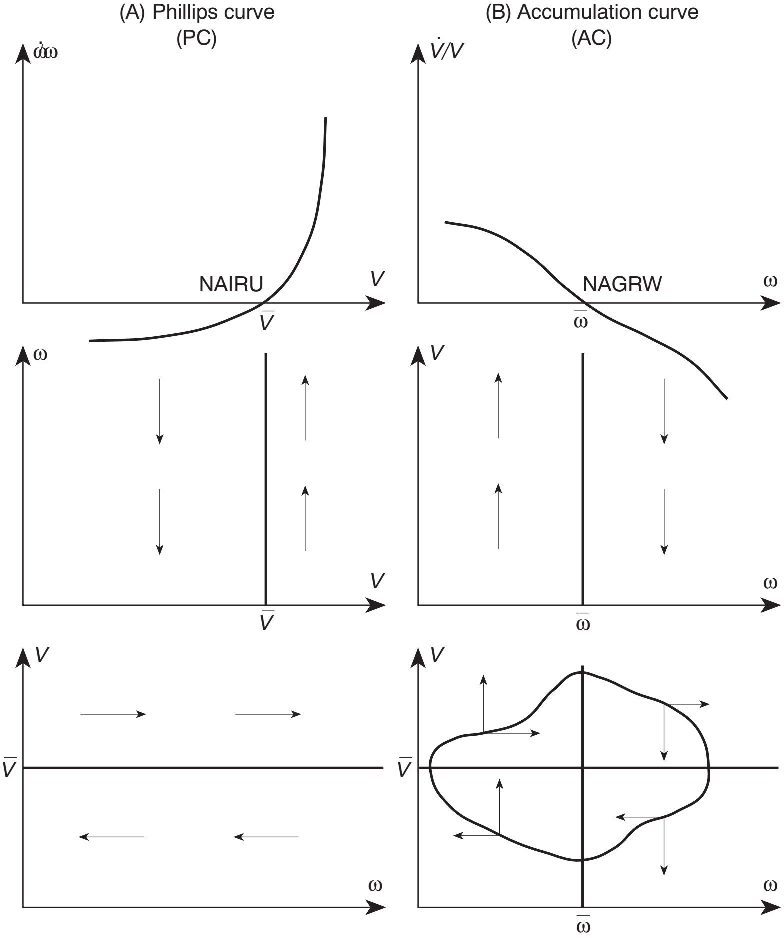


*Figure 6.2* The Goodwin (1967) growth cycle model

- there is a uniquely determined **N**on-**A**ccelerating-**I**nflation **R**ate of **U**tilization $\bar{V}$ of the labor force,[3]
- the PC curve exhibits negative values below this rate, and thus implies falling real wages in this domain,
- the PC shows positive values to the right of the NAIRU, implying rising real wages on this side.

We thus do not need in the following that the PC, in terms of employment, is monotonically increasing as is generally assumed in the literature.[4]

Corresponding to this real wage PC we assume as second building block of our model an Accumulation Curve (AC) which postulates that the rate of change

$\hat{V}$ of the rate of employment is a function of the level of real wages, again with a benchmark value, a **N**on-**A**ccelerating-**G**rowth **R**ate of **W**ages or NAGRW $\bar{\omega}$ of the real wage, which separates rising from falling rates of employment. As before the conclusions of the Goodwin growth cycle analysis do not demand the monotonicity of the curve shown top right in Figure 6.2.

On the basis of the AC and PC curves shown in Figure 6.2 one gets the adjustments of the rate of employment and of real wages as shown in the middle of this figure. In order to get from that the dynamic consequences for the interaction of real wages with the rate of employment (bottom right in Figure 6.2) one has to mirror the implications of the PC part of the model along the 45° degree line (bottom left). The phase space bottom right then simply integrates the neighboring situations as shown by the arrows in this space. The further implication of this model, that all curves (in the positive part of the phase space shown) must be closed orbits, can of course not be proven in this way. In order to get this result in an intuitively understandable way, one has to consider the following type of function:

$$L(\omega, V) = \int_{\bar{V}}^{V} PC(\tilde{V})/\tilde{V} \ \mathrm{d}\,\tilde{V} - \int_{\bar{\omega}}^{\omega} AC(\tilde{\omega})/\tilde{\omega} \ \mathrm{d}\,\tilde{\omega}.$$

The graph of this (Liapunov) function has the form of a global sink (under the assumptions made) with its minimum at $\bar{\omega}, \bar{V}$ and with all level curves (where the function assumes a given value) closed. Projected into the $\omega, V$ phase space these closed curves are just the orbits of the considered dynamics, since it is easily shown that $L$ is constant along the trajectories of the investigated dynamical system (Figure 6.2, bottom right). We add that this proof applies, on the one hand, to very general situations as far as functional shapes of the PC and AC curves are concerned, but that it, on the other hand, has to be checked carefully for (and will often not be applicable to) systems which do not rely on the simple cross-dual nature of the AC-PC interaction shown in Figure 6.2.[5] We thus have that all trajectories generated by the interaction of PC and AC dynamics represent periodic motions of the real wage and the rate of employment as the one shown in Figure 6.2, bottom right. We do not describe this overshooting dynamic in its details here, since this has been done many times already, including the original article of Goodwin (1967).

### *6.2.2 AC-PC model building: Extensions*

Desai (1973) has extended the Goodwin (1967) model by an explicit treatment of price level dynamics, through a delayed type of markup pricing.[6] Using in addition a money wage Phillips curve where price inflation enters additively with a factor $\eta$ that can be less, equal or larger than one, one can show by means of the above Liapunov function $L$ that the steady state $\bar{\omega}, \bar{V}$ of these extended dynamics becomes a global sink (a global source) if $\eta < 1$ ($\eta > 1$) holds true (while $\eta = 1$ gives the original Goodwin growth cycle). This is one among a variety of examples which shows that this growth cycle represents a border case between asymptotic stability and instability, of which one can thus say that its closed orbit structure is exceptional (structural unstable). Nevertheless, its message of overshooting conflicts about income distribution can be found in many advanced models that rely on a dynamic interaction of wages, prices and factor utilization rates.

It is thus one aim of the present chapter to show the power of this approach for disequilibrium growth theories of an integrated nature where also goods and asset market behavior is taken into account. Figure 6.3 indicates already what complexities this might add to the growth cycle just considered. Taking disequilibrium on the market for goods seriously in our view means that measures that represent this disequilibrium must be introduced into the analysis, which we shall do here by referring, on the one hand, to the rate of capacity utilization $U_c = Y/Y^p$ of firms and, on the other hand, to their rate of inventory disequilibrium $U_n = N/N^d$. Here, $Y$ denotes actual output and $Y^p$ potential output (for a fixed proportions technology), while $N$ denotes actual inventories and $N^d$ desired ones. Furthermore, we shall use in the following Figure 6.3 also the measures $V^w$, the (inside) employment rate of the employed workforce (based on over- or under-time work of the employed), and $\rho_l, r_l - \pi^e$ the rate of profit and the real rate of interest, $\pi^e$ the expected rate of inflation (everything here conceived as average over the longer run). Allowing for varying rates of capacity utilization within firms implies in addition that we now have to distinguish between the actual rate of employment, $V$, and the potential one, $V^p$, based on fully utilized capital stock and on a normal working-day for all members of the workforce.

Obviously, the AC dynamics now concerns the relationship between real wages and the growth rate of potential output, via the resulting profitability of firms and the investment decision based on them. Furthermore, rates of capacity utilization $U_c$ and the real financing costs of firms as measured by $r_l - \pi^e$ may also influence their investment decision and thus the growth path of potential employment and must therefore be added as 'shift' terms to the AC dynamics as indicated in Figure 6.3, top right. Note that we do not add the inventory measure here since we believe that this measure is related to short-run (pricing) decisions of firms solely. The PC dynamics top left, on the other hand, will be positively influenced by the internal employment rate of firms, and negatively (through positive changes in the price level) by the two rates that characterize the disequilibrium experienced by firms on the market for goods. We thus get that the two curves underlying the Goodwin growth cycle mechanism are neither fixed nor do they give rise to a unique phase space diagram, bottom right, as was the case in Figure 6.2. Instead, actual and potential employment will in general differ in a model that includes the interaction of goods and asset markets (the real–financial interaction), so that higher dimensional representations may become unavoidable when the analysis of these additions is approached.

The phase diagram bottom right in Figure 6.3 indicates in simple terms that the outcome of an integrated treatment of income distribution, goods market and asset market behavior may be very uncertain at the present stage of the investigation and may or may not lead to the result that Figure 6.2 will continue to play a prominent role in such extended dynamics. We will return to this topic in Section 6.5.

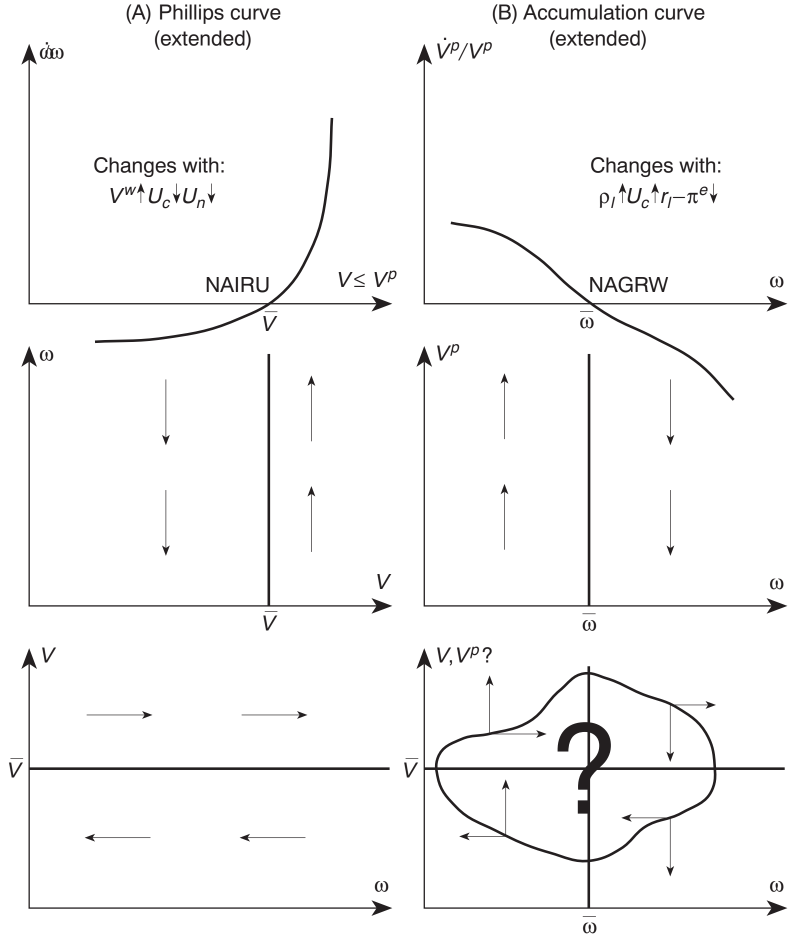


*Figure 6.3* The Goodwin growth cycle model in perspective (with influences from goods and asset markets)

We now investigate two fundamental extensions of the Goodwin growth cycle model that integrate, on the one hand, financial assets and liabilities, and, on the other hand, aspects of fluctuating goods market behavior in a still very basic way, in order to show how the dynamics are changed by such basic additions. We consider first the case where firms do not only use retained profits to finance their investment, but also loans (from asset holding households), while worker households still spend what they get. The following model is based on Keen (1999), investigated in detail in Chiarella, Flaschel and Semmler (2000), and is only briefly considered in the present chapter. For further details the reader is therefore referred to these other works.

Let us consider the budget equation of firms first. This equation simply states that the excess of nominal investment expenditures $pI$ over pure profits of firms $\rho pK$ is to be financed by new loans as shown in the following equation ($D$ the stock of loans of firms):

$$\dot{D} = pI - \rho pK.$$

This is here to be combined with a money wage PC of the type $\hat{w} = \beta_w(V - \bar{V})$, with $w$ the money wage, and an investment equation of the type $I/K = \alpha(\rho - \rho_{min})$ with $\alpha > 1$ and the definition of the pure rate of profits $\rho$ (net of interest payments $rD$ on loans) given by: $pY^p - rD - wl_y Y^p/pK$. The price level $p$ and the rate of interest $r$ are considered as given for the time being, as well as the benchmark rate of profit $\rho_{min}$. The parameter $l_y$ denotes the labor coefficient of the assumed fixed proportions technology (the other coefficient being given by the potential output–capital ratio $y^p = Y^p/K$). Since this is still a supply side dynamics, actual output will always equal potential output due to Say's Law assumed to prevail in this type of model.

These structural equations of the model give rise to:

$$\begin{aligned} \hat{u} &= \beta_w(V - \bar{V}), && \text{the reformulated PC dynamics} && (6.1)\\ \hat{V} &= \alpha(\rho - \rho_{min}) - n, && \text{the reformulated AC dynamics} && (6.2)\\ \dot{d} &= \alpha(\rho - \rho_{min})(1 - d) - \rho, && \text{the reformulated budget equation of firms} && (6.3) \end{aligned}$$

where $u = w/p \cdot l_y$ denotes the share of wages, where $n$ is the rate of natural growth, $d$ the debt to capital ratio $D/(pK)$ of firms, and $\rho = y^p(1 - u) - rd$ the pure rate of profit in this supply driven approach to economic growth. The dynamically endogenous variables are $u, V$ as in the original Goodwin model and a new one, the ratio $d$.

**Proposition**[7]

1. *Assume* $0 < r < n$. *Then: The steady state*

$$\begin{aligned} V_o &= \bar{V} + n/\beta_w && (6.4)\\ u_o &= (y^p - \rho_o - rd_o)/y^p, \quad \rho_o = \rho_{min} + n/\alpha && (6.5)\\ d_o &= 1 - \rho_o/n = (\alpha - 1)/\alpha - \rho_{min}/n && (6.6) \end{aligned}$$

*of the dynamics (6.1) – (6.3) is locally asymptotically stable.*

2. *This steady state is not globally asymptotically stable with respect to shocks of the debt capital ratio d which, when sufficiently large, can lead to an explosive development of the debt ratio d.*

The details of this proposition and its proof are provided in Chiarella, Flaschel and Semmler (2000) and will not be repeated here. We simply conclude here that debt financed investment makes the Goodwin growth cycle convergent for small

$r$ and for small shocks in its state variable $d$, but that sufficiently high debt per unit of capital can make these dynamics divergent ones, implying a situation of corridor stability in the place of the closed orbit structure of the original Goodwin (1967) approach.

### 6.2.3 Debt deflation and adverse real wage adjustments

It is of course necessary to add a theory of the price level to the above model of debt accumulation, due to its long-run nature. We thus now integrate goods market disequilibrium and fluctuating rates of capacity utilization. This will lead us to a framework that now exhibits two interacting processes of income distribution, between profits and wages on the one hand and between profits and interest on the other hand. We thus now extend the model (6.1) – (6.3) to include in it the possibility for price level deflation and thus the possibility for the occurrence of debt deflation (high levels of debt combined with declining profitability due to falling output prices) as well as the possibility of an adverse adjustment of real wages, both leading to instability of the steady state of the model. This gives rise to the following nominal dynamics for wages $w$, prices $p$ and the debt ratio $d$ coupled with an investment driven growth and employment path, the details of which are explained below:

$$\begin{aligned}
\hat{w} &= \frac{1}{1-\kappa_w\kappa_p}[\beta_w(V-\bar{V})+\kappa_w\beta_p(U_c-\bar{U}_c)](+\pi^e) && (6.7)\\
\hat{p} &= \frac{1}{1-\kappa_w\kappa_p}[\kappa_p\beta_w(V-\bar{V})+\beta_p(U_c-\bar{U}_c)](+\pi^e) && (6.8)\\
\hat{V}^p &= \alpha(\rho-\rho_{min})-n && (6.9)\\
\dot{d} &= \alpha(\rho-\rho_{min})(1-d)-\rho-\hat{p}d && (6.10)
\end{aligned}$$

These dynamics are based on the following static (and linearized) relationship representing Keynesian goods market equilibrium, here directly expressed in terms of the rate of capacity utilization $U_c$ of firms:

$$U_c = \bar{U} + d_1(\frac{w}{p} - (\frac{w}{p})_o) + d_2(d-d_o), \qquad d_1, d_2 \leq 0$$

This equation is used here as a shortcut for the delayed feedback chain on the market for goods to be introduced in Section 6.5 on the KMG model type (and its richer concept of aggregate demand). We have assumed in this equation that output and capacity utilization depend negatively on the real wage, based on the particular view that the negative real wage effect on investment dominates the positive one on consumption (the orthodox point of view), and have also assumed that output and capacity utilization depend negatively on the debt ratio $d$, again because investment depends negatively on it. The above goods market representation allows for Rose (1967) type real wage effects of traditional type (where price flexibility will be destabilizing) and for Fisher debt effects (where price flexibility will also be destabilizing), but it still excludes Mundell-effects, for example based on the inflationary expectations mechanism considered in Section 6.3. Finally we have $\rho = y - \frac{w}{p} l_y y - rd = y^p U_c (1 - \frac{w}{p} l_y) - rd$ for the rate of pure profits $\rho$.

The first two laws of motion for wages $w$ and prices $p$ can be easily derived, under one additional assumption stated below, from the following wage–price adjustment equations:

$$\hat{w} = \beta_w (V - \bar{V}) + \kappa_w \hat{p} + (1 - \kappa_w) \pi^e, \quad 0 < \kappa_w < 1 \tag{6.11}$$

$$\hat{p} = \beta_p (U_c - \bar{U}_c) + \kappa_p \hat{w} + (1 - \kappa_p) \pi^e, \quad 0 < \kappa_p < 1 \tag{6.12}$$

These equations represent two symmetrically formulated Phillips curves in the place of the hybrid single one that is usually considered in the literature, see Solow and Stiglitz (1968) for an early formulation of this type, Rose (1990) for a recent and more advanced one, and Fair (1997a,b) for an application of such an approach. These two equations state that wage as well as price inflation depend positively on the demand pressure in the market for labor or goods, respectively, and on a weighted average of the relevant cost-push expression for each of these PCs, actual price inflation in the first and actual wage inflation in the second case both combined with an average rate of inflation $\pi^e$ that is expected to hold over the medium run. This rate is set equal to zero in the present section for reasons of simplicity. Investigation of the role of expected inflation will be started with Section 6.3 and will be fully present in Section 6.5 on the KMG model.

The equations (6.7), (6.8) are easily interpreted and they state that wage as well as price dynamics can be expressed in terms of both demand pressure variables solely, by appropriate elimination of the cost-push expressions originally contained in them. Increasing demand pressure in one of these markets is therefore already sufficient to raise both wage and price inflation rates. Note in this regard also that there is a second **N**on-**A**ccelerating-**I**nflation **R**ate of **U**tilization $\bar{U}_c$ now present in the model, for the goods market and the rate of capacity utilization of the capital stock, which plays a similar benchmark role for price inflation as the rate $\bar{V}$ did for wage inflation. Demand pressure is therefore always measured relative to such benchmark rates, both assumed to be less than 1.

The other two equations (6.9), (6.10) are the same as before, with the exception that $\hat{p}d$ has now to be added to (6.10), due to the definition $d = D/(pK)$ of the ratio $d$, since the price level $p$ is now a variable of the model. For simplicity we assume in the following that the minimum rate of profit $\rho_{min}$ of investors is equal to $r$, the given rate of interest of the model.[8]

Note finally that the equation for the actual rate of employment $V$ is related to the rate of capacity utilization $U_c$ in the following way:

$$V = l_y Y/L = l_y (Y/Y^p)(Y^p/K)/(L/K) = l_y U_c y^p / l, \quad L \text{ the supply of labor, } \hat{L} = n$$

where $l_y, y^p$ are again the labor coefficient and the potential output–capital ratio of the fixed proportions technology and where $l = L/K$ denotes the factor endowment ratio of the economy. Note also that the third equation (6.9) now concerns the evolution of the potential rate of employment $V^p = l_y Y^p / L$ as discussed in connection with Figure 6.3. The dynamical system (6.7) – (6.10) therefore needs further reformulation in order to make it an autonomous system of dimension 4, since $V$ now depends on $U_c, l$ and since (6.9) can no longer be used to describe the actual evolution of the rate of employment as in the earlier treatments of the Goodwin approach to cyclical growth (where there was always full capacity growth).

A simple reformulation of the dynamics (6.7) – (6.10) is in this regard provided by making use of the following relationship between the ratios $V, V^p, U_c$ :

$$V = l_y Y/L = l_y Y^p/L \cdot Y/Y^p = V^p \cdot U_c$$

by which the rate of employment can be removed from the above 4D dynamics which are then based on the state variables $w, p, V^p, d$, since $U_c$ has been assumed to be a function of $w, p, d$.

The interior steady state of these dynamics in the state variables $w, p, V^p, d$ is characterized by:[9]

$$\begin{aligned}
d_o &= 1 - r/n && (6.13)\\
U_c^o &= \bar{U}_c && (6.14)\\
V_o^p &= U_c^o/\bar{V} && (6.15)\\
\rho_o &= y^p U_c(1-\omega_o l_y) - rd_o = r, \quad \omega = w/p && (6.16)\\
\omega_o &= \frac{1-(\rho_o + rd_o)/(U_c y^p)}{l_y} && (6.17)\\
p_o &= \text{determined by initial conditions} && (6.18)\\
w_o &= p_o\omega_o && (6.19)
\end{aligned}$$

It is therefore in fact not uniquely determined as far as nominal magnitudes are concerned, since these dynamics can be further reduced to an autonomous system in the real variables $\omega, V^p, d$ due to the fact that equations (6.7), (6.8) can be transformed into the single law of motion for the real wage:

$$\hat{\omega} = \frac{1}{1-\kappa_w\kappa_p}[(1-\kappa_p)\beta_w(V-\bar{V}) - (1-\kappa_w)\beta_p(U_c - \bar{U}_c)] \tag{6.20}$$

All dynamical equations (6.7) – (6.10) therefore only depend on $\omega, V^p, d$ which means that the 4D system has a singular Jacobian at the steady state and thus exhibits zero root hysteresis with respect to the nominal variables of the model which are thus determined in their long-run behavior by historical conditions. The law of motion for nominal prices (and wages) thus can be treated as appended to the real dynamics.

**Proposition**

1. *Assume $0 < r < n$ and that $\beta_p, \kappa_p, d_2$ are all sufficiently small. Assume furthermore that the investment parameter $\alpha$ is such that $\alpha r - n > 0$ holds true. Then: The steady state (6.13) – (6.19) of the dynamics (6.7) – (6.10) is locally asymptotically stable.*

2. *The steady state (6.13) – (6.19) of the dynamics (6.7) – (6.10) is not locally asymptotically stable for all price adjustment speeds* $\beta_p$ *chosen sufficiently large.*

3. *Assume that nominal wages are completely fixed* $(\beta_w, \kappa_w = 0)$. *Then: The dynamics (6.7) – (6.10) is monotonically explosive, implying higher and higher real wages and debt to capital ratios, for all initial debt capital ratios sufficiently high and all real wage levels above their steady state value.*

**Proof:** See Chiarella, Flaschel and Semmler (2000).

Sufficiently sluggish price level adjustments are thus favorable for local asymptotic stability, while sufficiently flexible price levels will definitely destroy it. This is due to the joint working of destabilizing Rose (1967) real wage and Fisher (1933) debt deflation mechanisms. On the one hand, if prices are more flexible than wages we see that depressions will increase the real wage, since prices fall faster than nominal wages in such a situation, deepening the depression already under way, an adverse Rose-effect as in Rose (1967), but here no longer in a framework of Keynes–Wicksell type. On the other hand, if prices are sufficiently flexible, downward in the case of a deflationary situation, they will raise real debt $d$ significantly, see (6.10). This depresses economic activity further and thus leads to a deflationary spiral as was happening to some extent during the Great Depression of the 1930s and as has again to some extent been the fear of policy makers in the years 1998/9 and at present.

We thus end up with a model type and its implications that has extended AC-PC analysis considerably, with regard to the range of PC-dynamics to be used and with respect to Keynesian demand pressure appearing in PC as well as AC-dynamics now. Nevertheless, the PC discussion as well as the treatment of Keynesian demand problems must be further improved in the light of what has been shown, which will be done in Sections 6.4 and 6.5, thereby continuing the discussion of the question mark in Figure 6.3 of this section. In closing this section we remark that the Goodwin (1967) growth cycle model has of course been extended in numerous other ways after its publication which cannot be surveyed here due to lack of space. In this regard the reader is referred to Flaschel, Franke and Semmler (1997), Chiarella and Flaschel (2000b), where also the question of global boundedness of locally diverging dynamics is pursued, see also Section 6.5 of the present chapter.

## 6.3 IS–LM–PC analysis

We have started the analysis of disequilibrium growth in the preceding section from the supply side and from a prototype growth cycle model whose relevance for integrated disequilibrium growth still remains to be investigated. However, we have augmented these dynamics by a rudimentary theory of aggregate demand already, and have seen that this introduces significant destabilizing feedback mechanisms into these supply side dynamics, the real wage Rose-effect and the Fisher debt-effect, when the price level is assumed as sufficiently flexible. In this section we will now start the analysis from the opposite end, from the demand side, by means of the conventional IS–LM model, augmented by PC dynamics which includes an adaptive formation of expectations, as has often been discussed in the literature. We will argue in this section that this well-known dynamic model should be well-understood meanwhile, but that indeed just the opposite will turn out to be true. This section will therefore reveal, on the one hand, the true power of the IS–LM–PC approach (giving rise to persistent fluctuations in place of the generally assumed global asymptotic stability of the steady state of the model). It will, on the other hand, show that this approach is at least as limited with respect to a full understanding of the fundamental feedback mechanisms of macrodynamic systems, discussed in the introduction to this chapter, as the basic AC-PC dynamics considered in Section 6.2. Both approaches are in fact complementary to each other, since IS–LM–PC analysis introduces interest rate flexibility (and the Keynes-effect that is based on it), inflationary expectations (and the Mundell-effect that derives from them). We will study their interaction with respect to the stability question once again. Sections 6.2 and 6.3 will be integrated with each other in Section 6.5 of this chapter which will move us closer to the perspective introduced in Section 6.1.

### *6.3.1 Medium-run IS–LM analysis?*

Macroeconomic textbooks (also on the advanced level) usually include sections which extend the IS–LM model to the medium run, where wage and price adjustment occur depending on the state of the labor market, where then expectations and NAIRU augmented PC dynamics are employed in order to show or indicate the stability of the ‘full employment’ or NAIRU-equilibrium. This is in particular true for the prominent textbook by Dornbusch and Fischer (1996) who make use of the following simple IS–LM–PC dynamics[10] in order to discuss on this basis supply side and demand side shocks and the subsequent readjustments to the NAIRU rate of employment:[11]

$$\begin{aligned} \dot{Y} &= a_1(\bar{\mu} - \pi) + a_0\bar{f}, \quad \hat{M} = \bar{\mu} = const \\ \pi &= \hat{p} = \hat{w} = \beta_w(Y - \bar{Y}) + \pi^e \\ \dot{\pi}^e &= \beta_{\pi^e}(\widehat{p} - \pi^e) \end{aligned}$$

This model is based on a dynamic theory of effective demand whereby the time rate of change $\dot{Y}$ of IS–LM equilibrium output $Y$ is postulated to depend positively on the rate of change of real balances $M/p$:

$$\widehat{M/p} = \hat{M} - \hat{p} = \bar{\mu} - \pi$$

(due to the conventional Keynes-effect of static IS–LM theory) and on an exogenously given *dynamic* fiscal policy parameter $\bar{f}$. The next equation then adds a linear, expectations augmented, natural rate (money wage and price level) PC here based on output levels in the place of rates of unemployment. Since this model is based on fixed proportions in production, on a constant labor supply and on a constant markup on average wage costs, this equation can however easily be translated back into one that shows rates of unemployment (or employment) in the place of $Y$, see below. Furthermore, the assumption on markup-pricing immediately implies that wage and price inflation can be identified and represented by a unique magnitude $\pi$.

The third equation finally is the conventional adaptive expectations mechanism of elementary inertia theories of inflation and stagflation ($\pi^e$ the expected rate of inflation). The above model can be reduced to the following form:

$$\begin{aligned}\dot{Y} &= a_1(\bar{\mu}-\pi)+a_0\bar{f}\\ \dot{\pi} &= \beta_w a_1(\bar{\mu}-\pi)+\beta_{\pi^e}\beta_w(Y-\bar{Y})+\beta_w a_0\bar{f}\end{aligned}$$

which is a linear autonomous differential equations system of dimension 2 in the variables output $Y$ and inflation $\pi$.

These dynamics imply everything one would like to find in a basic model of monetarist wage–price dynamics with adaptive expectations, here however in the context of a system that is apparently of IS–LM–PC type. There is a unique and economically meaningful steady state $Y_o = \bar{Y}, \pi_o = \bar{\mu} + a_0\bar{f}/a_1$ which reduces to $Y_o = \bar{Y}, \pi_o = \bar{\mu}$ if fiscal policy is stationary. This steady state is globally asymptotically stable in the whole phase plane for all possible parameter values of the model and is surrounded by cyclical forces when adjustment of inflationary expectations is fast. There hold the monetarist propositions on monetary policy, accelerating inflation, periods of inflation and stagflation, long-run neutrality, changing expectations mechanisms in this framework of medium run IS–LM dynamics hold. A detailed discussion of all this – which due to the linearity of the model is straightforward – is provided in Dornbusch and Fischer (1996) and Flaschel and Groh (1996, Ch.4, 1998) and will not be repeated here, since we shall argue now that this model of Dornbusch and Fischer (1996, Ch.16) is not a valid extension of their linear IS–LM analysis[12] (which we do not question) towards an inclusion of the dynamics of wages, prices and inflationary expectations.[13]

Two simple observations must here suffice to justify the claim that this type of analysis is an invalid one even when based on the assumptions made in Dornbusch and Fischer (1996). The first observation is that investment depends on the real rate of interest in their book (rising inflationary expectations $\pi^e$ will move the IS-curve to the right and thus must enter the original $\dot{Y}$ equation by mathematical necessity). The second observation is that the Phillips curve is a nonlinear dynamic equation, since it is based on the growth rate and not the time derivative of money wages. However transformed, the dynamics to be analyzed is thus always nonlinear and thus cannot be represented in the large by the above linear dynamics. The above dynamics therefore do not represent a correct formalization of the Dornbusch and Fischer (1996) assumptions about the Keynesian short- and the monetarist medium-run and thus should be dismissed from this book for these and other reasons.

### *6.3.2 IS–LM–PC analysis proper*

The first correction of the dynamics on the basis of the assumptions of Dornbusch and Fischer's (1996) book, is that the outcome of their linear IS–LM model, reformulated in terms of the employment rate $V = (Y/x)/L$, should be represented as follows

$$V = a_0 + a_1 m + a_2 \pi^e, \quad a_1, a_2 > 0 \tag{6.21}$$

An increase in real balances $m = M/p$ will increase IS–LM equilibrium output (the Keynes-effect) and an increase in inflationary expectations $\pi^e$ will do the same (the so-called Mundell effect), due to the rightward shift of the IS curve that results from this parametric change. The PC dynamics, easily transformed to rates of change $\hat{m}$ of real balances $m$ for a given growth rate $\bar{\mu}$ of the money supply, then read (when output data are transformed to rate of employment expressions):

$$\widehat{m} = \bar{\mu} - \beta_w (V - \bar{V}) - \pi^e \tag{6.22}$$

$$\dot{\pi}^e = \beta_{\pi^e}(\pi - \pi^e) = \beta_{\pi^e}\beta_w (V - \bar{V}) \tag{6.23}$$

There is of course again the assumption of adaptive expectations in order to make the model determinate. This is the complete model of the Dornbusch and Fischer (1996) approach to medium run wage–price dynamics. We shall see that this proper IS–LM–PC dynamics has little in common with the dynamics considered in the preceding subsection, which at least disqualifies some of the monetarist conclusions there stated.İongpage

**Proposition**

*There are always two steady states of the dynamics (6.22) – (6.23), one that is interior to the right half of the phase plane and thus economically meaningful and one that lies on its boundary:*

$$m_0 = (\bar{V} - a_0 - a_2\bar{\mu})/a_1 > 0, \; \pi_0^e = \bar{\mu} \quad \textit{and} \quad m_0 = 0, \; \pi_0^e = (\bar{V} - a_0)/a_2 > 0.$$

*The dynamics around the border steady state are always of saddlepath type (*$\det J < 0$*), while the dynamics around the interior steady state will be represented by a stable node, a stable focus, an unstable focus and an unstable node as the parameter* $\beta_{\pi^e}$*, the adjustment speed of inflationary expectations, is increased from close to zero to close to infinity.*

The proof of these results is simple and is provided in Flaschel and Groh (1998). These results in particular state that the dynamics are *never* globally asymptotically stable and are also not locally asymptotically stable if inflationary expectations are adjusted with sufficient speed. The monetarist belief in the overall asymptotic stability of the private sector is therefore not justified in an IS–LM–PC framework, which at best allows for corridor stability (when the Keynes-effect is sufficiently strong relative to the Mundell-effect), but not for more, see also Groth (1993) on this matter. If the steady state is locally explosive it will be globally explosive. The dynamics therefore are not a viable one in this case and the question must be posed as to what can make them bounded in such an explosive situation.

Keynes (1936) in fact did provide the basic answer to this question when stating:

> Thus it is fortunate that workers, though unconsciously, are instinctively more reasonable economists than the classical school, inasmuch as they resist reductions of money-wages, which are seldom or never of an all-round character . . . (p.14)
> The chief result of this policy (of flexible wages, P.F.) would be to cause a great instability of prices, so violent perhaps as to make business calculations futile . . . (p.269)

We use the following stylized modification of the PC used so far in order to provide a mathematical expression for the institutional fact just quoted:

$$\widehat{w} = \max\{\beta_w(V - \bar{V}) + \pi^e, 0\}.$$

This Phillips curve says that money wages behave as in the preceding subsection if their growth rate is positive, but stay constant if they would be falling in the previous situations. There is thus no wage deflation possible now. This assumed kink in the money wage PC could be smoothed or some wage deflation could be allowed for, but this will not alter the conclusions significantly. We consider this kinked Phillips curve as a much better description of reality than the one that is linear throughout.[14]

The immediate consequence of this new form of the Phillips curve is that system (6.22) – (6.23) now only applies when $\beta_w(V - \bar{V}) + \pi^e \geq 0$ holds while it must be replaced by

$$\widehat{m} = \bar{\mu} \tag{6.24}$$

$$\dot{\pi}^e = -\beta_{\pi^e}\pi^e \tag{6.25}$$

in the case $\beta_w(V - \bar{V}) + \pi^e < 0$.[15] We thus get a system of differential equations which is only continuous now, but which can be made a smooth system in an obvious way. We call this system the patched system while we refer to the earlier dynamics as the unpatched one.

There are a variety of propositions that can be formulated in the context of such a patched dynamics, see Flaschel and Groh (1998), but due to space limitations we will consider here only one of them which describes the outcome of the explosive case of the Dornbusch and Fischer (1996) model in the patched situation when there is steady state inflation ($\bar{\mu} > 0$). This proposition refers to Figure 6.4 which represents the considered dynamics in the $m, \pi^e$ state space.

**Proposition (Viability Theorem):**

1. *There exist exactly three steady states for the patched dynamics: $S_o, S_1, W$ if $\bar{\mu} > 0$ holds. These steady states are connected by the $\dot{\pi}^e = 0$ isocline.*

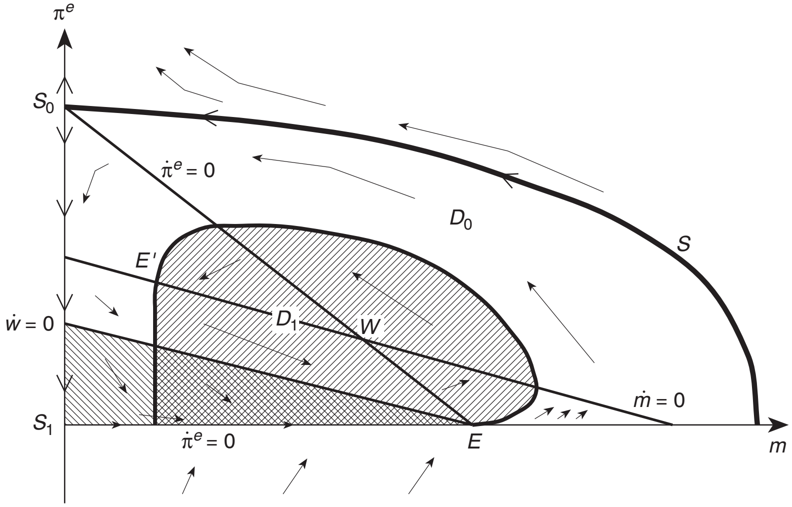


*Figure 6.4* Implications of the kinked Phillips curve in the case of steady state inflation

2. *Assume that the interior steady state W is locally repelling (for values of $\beta_{\pi^e}$ chosen sufficiently large). Then: every trajectory in $D_o$ converges to a persistent cycle around W (and in $D_1$).*

As this figure shows the domain below the separatrix $S$ of the saddlepoint $S_0$ in the nonnegative orthant is now an invariant domain $D_o$, i.e. no trajectory which starts in it can leave it. Note also that the domain below the $\dot{w} = 0$ isocline is governed by the above revised dynamics in place of the one of the preceding section, which however only alters the direction of the dynamics on the horizontal axis. This axis is now also an isocline ($\dot{\pi}^e = 0$) of the patched dynamics (up to point $E$). Note furthermore that the trajectory which starts in $E$, followed up to point $E'$ and then vertically continued up to the m-axis, also defines an invariant domain $D_1$ of the patched dynamics which moreover is attracting for all trajectories in the interior of $D_o$. We thus have that all orbits in $D_o$ (with the exception of the ones on the vertical axis) are either inside of $D_1$ or are entering the domain $D_1$ (from its left) at some point in time.

We do not go into the proof of this proposition or a deeper explanation of the phase diagram shown, but refer the reader to the quoted work in this regard. We simply close this subsection here by stating that the situation in Figure 6.4 is far away from anything that can be found on IS–LM–PC dynamics in the literature, and this is simply due to the fact that there is a Mundell effect (in IS–LM), a simple growth law (of wages) and a fundamental institutional asymmetry in the PC analysis to be used, which makes the overall dynamics viable up to shocks in $m$ or $\pi^e$ that go beyond the shown separatrix $S$.

### *6.3.3 Reinterpreting the IS–LM–PC dynamics*

We now proceed to a reformulation and reinterpretation of the structural equations underlying the dynamical system of the preceding subsection which do not alter its mathematical formulation and stability features, but which give the goods market in the place of the labor market the decisive role in the explanation of the stability and instability scenarios just discussed. In order to introduce this reformulation of IS–LM–PC analysis, we start from a stylized representation of empirical results (testing the conventional NAIRU model) provided by Fair (1997a), and extended in Fair (1997b). Fair's (1997a) reconsideration of the structural price and wage equation leads him to the (here simplified) result that it is in fact the price Phillips curve which determines the shape of the integrated Phillips curve of the literature, while wage inflation is following price inflation more or less passively. This implies that the integrated Phillips curve now refers to demand pressure on the goods market and not as is customarily assumed on the labor market. In terms of the two wage and price Phillips curves considered in the preceding section this leads to their following special reformulation

$$\hat{w} = \kappa_w \hat{p} + (1-\kappa_w)\pi^e \tag{6.26}$$
$$\hat{p} = \beta_p(U_c - \bar{U}_c) + \kappa_p \hat{w} + (1-\kappa_p)\pi^e \tag{6.27}$$

or, if solved as in the preceding section, but now with expected medium-run inflation shown explicitly (no longer equal to zero):

$$\hat{p} = \frac{\kappa_p}{1-\kappa_w\kappa_p}\beta_p(U_c - \bar{U}_c) + \pi^e. \tag{6.28}$$

This is the same type of Phillips curve as used in the preceding subsection with the only (economically seen very important) difference that the rate of capacity utilization is now used in place of the rate of employment as the measure of demand pressure that drives price inflation. Formally seen, this Phillips curve can even be represented exactly as in the preceding subsection (in the situation where labor supply and capital stock growth are still excluded due to the medium run nature of the performed analysis), if one makes use of the following implications of our assumption of a fixed proportions technology:

$$U_c = \frac{Y}{Y^p} = \frac{1}{l_y y^p}\frac{L^d}{K} = \frac{1}{l_y y^p}\frac{L}{K}\frac{L^d}{L} = \frac{1}{l_y y^p}\frac{L}{K}V$$

The price Phillips curve thereby becomes of the form:

$$\hat{p} = \tilde{\beta}_p(c_1 \cdot V - c_0) + \pi^e. \tag{6.29}$$

without implying that the demand pressure driving inflation is a labor market phenomenon and also not that the NAIRU rate $c_0/c_1$ implied by it is related to labor market issues. In fact the above shows that it is the NAIRU rate of firms' capacity utilization while there is in fact no labor market NAIRU at work in the present model. This shows that there is the possibility that the literature on the conventional type of integrated PCs has completely misinterpreted the NAIRU phenomenon.

On the basis of this reformulation of the (across markets) integrated Phillips curve the analysis of the preceding subsection can be repeated word by word, with the interpretational differences just stressed, namely that $\bar{U}_c$ determines the steady state value of the rate of employment and that the destabilizing Mundell-effect is now basically due to the behavior of firms. This implies that price flexibility is bad for economic stability in a third way (adding to the Rose adverse real wage effect and the Fisher debt effect considered in the preceding section), namely through the so-called Mundell effect, which says that there can be a positive (destabilizing) feedback mechanism leading from rising inflation to rising expected inflation and then, via the real rate of interest, to rising aggregate demand and thus to further inflationary impulses. This holds if the negative Keynes effect on aggregate demand based on a positive correlation of the nominal price level and the nominal interest rate is not strong enough to overthrow this cumulative tendency in the interaction of expected and actual rates of price inflation. Price flexibility must therefore be regarded with suspicion from at least three different angles, in particular when it occurs under deflationary pressure, since floors to economic activity are not so easily established as ceilings (the latter are built into the system via supply bottlenecks or via monetary and fiscal policy, which may stop accelerating growth, but which cannot so easily revive declining economic activity).[16]

A recent discussion concerning the core of practical macroeconomics, in the papers and proceedings issue of the *American Economic Review* 1997, provides numerous statements for and against the scope and relevance for traditional Keynesian dynamics, in particular in the applied area, with those against generally referring to a lack of microfoundations of Keynesian analyses, contrasted to the progress of the macroeconomic theory of the last two decades. Nevertheless, there continue to exist observations of the kind:

> Right or wrong, the IS–LM model, and its intellectual cousins, the Mundell-Fleming model and the various incarnations of aggregate supply – aggregate demand models, have proved incredibly useful at analyzing fluctuations and the effects of policy.
>
> (Blanchard, 1997, p.245)

Yet, in view of the analysis presented in this subsection and in the preceding one, it can be claimed that traditional Keynesian analysis is in fact still poorly understood (or at least represented) even on the textbook level, but also in more advanced types of analysis, if dynamic issues are addressed. There is no thorough discussion of the many scenarios even the simple IS–LM–PC model can give rise to, there is no investigation of the additional instabilities arising from Rose real wage and Fisher debt effects considered in Section 6.2, nor is there any far-reaching analysis of the full picture of traditional Keynesian dynamics as sketched in the introduction to this chapter and as filled with more details in the sections that will follow. There is no well-documented general analysis of the wage–price spiral, based on demand pressure and cost-push terms, as we shall present it in the following section, no detailed study of an integrated theoretical model with sluggish price as well as quantity adjustments and varying rates of capacity utilization for both labor and capital as we shall present in Section 6.5 and no such model where endogenous growth, financial markets, policy feedback rules, and more, are systematically investigated as to their contribution to the overall behavior of the macroeconomy.

There are however integrated models which address many of these issues, but not all of them, from an applied perspective. Yet, these empirically motivated structural macroeconometric models have until recently[17] never been analyzed in detail from the theoretical perspective, since these models are filled with a lot of empirical details, often not complete with respect to long run aspects, and not represented in the form of a theoretical reference model in the literature which would have allowed the analysis of their steady states, their stability and of the mechanisms that would ensure global boundedness should the steady state be surrounded by centrifugal forces. Recent structural macroeconometric models tend to include long run considerations now, but they still continue to believe that the deterministic part of the dynamics they consider is behaving like a shock absorber, a dynamic scenario with a long tradition in dynamic economic theory, but nevertheless only one possibility of many others, in particular in the high order dynamical systems any integrated macrodynamics will necessarily lead us to.

It can therefore not really be claimed that we do not have models at our disposal which enrich the early Keynesian quantity dynamics by detailed price dynamics, asset market behavior, questions of open economies and more, but it is surely true that these models or better their common theoretical core (removing lags, special features and pure replication by disaggregation) are not presented and investigated to any satisfying degree. In view of the above quotation from Blanchard (1997) it must therefore be added that the usefulness of dynamic IS–LM analysis is to be regarded as very limited and fragmented in its present state where there are only more or less isolated examples for such an analysis available, see Turnovsky (1977, 1995) for some left over ruins of this type. Furthermore, there is still the confusion, see again the above quotation from Blanchard (1997), that AS–AD dynamics, as for example presented in Sargent (1987, Ch.5), is the model of Keynesian medium and long run analysis. This, however, cannot be true simply due to the fact that capital is always fully employed in these approaches, while Keynesian dynamics should in principle study the reasons for the possible under-employment (or over-employment) of all factors of production.

We thus conclude that even those authors who show some sympathy for the traditional Keynesian way of analyzing the macroeconomy do not really describe what such analyses have been capable of solving and where they are still in their state of infancy (as far as systematic explorations of the fluctuating growth patterns they can give rise to are concerned). A step closer to the true alternative in macrodynamic analysis is Barro (1994) when he states:

> We have available, at this time, two types of internally-consistent models that allow for cyclical interactions between monetary and real variables. The con-

> ventional IS/LM model achieves this interaction by assuming that the price level and the nominal wage rate are typically too high and adjust only gradually toward their market-clearing values. The market-clearing models with incomplete information get this interaction by assuming that people have imperfect knowledge about the general price level.
>
> (Barro, 1994, p.4)

It is clearly stated here that both wage and price rigidity and imbalance in the labor and the goods market are the basic building blocks of the disequilibrium approach to macrodynamics, which cannot therefore be of the AS–AD variety. Yet it is not made clear that the analysis of the ‘cyclical interactions between monetary and real variables’ is not yet very far developed if the scenario presented in the introduction of this chapter (and more) is really taken seriously. However, it is admitted by Barro that disequilibrium approaches can be internally consistent.

We conclude this section with the observation that much remains to be done even on the level of traditional Keynesian IS–LM growth dynamics in order to obtain a well-understood reference situation against which the achievements of more recent studies of the dynamic implications of market imperfections, supply side bottlenecks (and also of perfect market clearing approaches) can be evaluated and put into perspective. To demonstrate this in more detail will be the topic of the remaining sections of this chapter, see also Chiarella, Flaschel, Groh and Semmler (2000).

In Section 6.4 we shall start this discussion with a critical evaluation of the generality of applied Phillips curve analyses where, as we shall see, numerous alternatives have been proposed and claimed to be the crucial ones, but where no unifying approach so far exists which conceives all these studies as special cases of a general formulation of wage price dynamics based on non-market clearing in both labor and goods markets. We shall supply in the section such a general approach as an extension of what we formulated in Section 6.2 and as a detailed representation of the supply side features of traditional Keynesian theory as summarized in Figure 6.1 of Section 6.1. On the basis of this general framework for wage price interactions we shall then add to the discussion of Section 6.3 a model (and theoretical interpretation) of the presented IS–LM–PC analysis of this section which further contributes to the insight gained, which are, that there is in fact not a uniquely determined understanding of medium-run IS–LM analysis. The topics to be solved by the IS–LM–PC approach are thus far from being settled and understood, in contrast to what is generally declared to be the case in the literature, where these issues are generally considered as sufficiently treated and on the basis of this understanding as outdated.

## 6.4 The two PCs approach: extensions and applications

We have already stressed the importance of the usage of separate Phillips curves for wage and price dynamics at several places in this chapter, in contrast to many statements on ‘the Phillips curve’ that can be found in the literature on macroeconomic theory. Nevertheless, there exists a long (basically off-stream) tradition to make use of two such curves in the economic theorizing, in particular in the growth cycle literature. We have already pointed in Section 6.2 to the work of Rose (1967), who uses two PCs with demand pressure terms solely and to Desai (1973) where an augmented wage PC is coupled with a delayed price PC of the cost-push variety. There is the early article of Solow and Stiglitz (1968) where symmetrically formulated wage and price PCs are used (both with demand pressure and cost-push terms) to investigate medium run dynamics where regime switching occurs, and there is the related macroeconomic literature of non-Walrasian type (Malinvaud (1980), Benassy, Picard, Hénin, Michel and others) where such PCs have often been used in conjunction with both labor and goods market disequilibrium, see Malinvaud (1980) for a typical example. Rowthorn (1980) makes use of a dynamic price PC coupled with a static wage PC in order to show how the conflict over income distribution allows for an endogenous determination of the NAIRU rate of capacity utilization of both labor and capital, an approach that will be briefly reconsidered in this section. There is finally recent work by Rose (1990) where PCs of the type (6.11), (6.12) were introduced and discussed. With respect to applied work on Phillips curves, we have already considered in simplified form the approach of Fair (1997a,b) in Section 6.3. Fair's (1997a) approach exhibits of course further arguments in the structural wage and price equations he proposes, see his page 6, such as labor productivity growth and import prices.

Specifying wage and price dynamics as two separate equations is highly desirable in theoretical as well as applied macroeconomic analyses since it will make explicit the reasons that may lead us to a single integrated Phillips curve later on, and since demand pressure variables should be specific to the price variable to considered and only be substituted by measures referring to other markets if there is good reason to do so. The scene is thereby set to further consider the two PC approach with demand as well as cost pressure terms from the theoretical as well as the applied perspective and to investigate its true degree of generality before special restrictions are added as far as the perspective of theory based macroeconometric model building is concerned.

As already stressed, Rose (1990) has revived the theoretical consideration of the two PCs approach, an approach taken up in Chiarella and Flaschel (2000b) in their formulation of the wage–price module of a hierarchically structured series of integrated models of monetary growth. In this work, however, the degree of generality chosen basically remained limited to the PCs (6.11), (6.12) as presented in Section 6.2 of this chapter. Yet, these two equations are but the beginning of truly general PC formulations and investigations (leaving aside here that PC findings are not limited to the macro-markets for goods and labor, but may also apply to sectoral macro-considerations concerning housing, agriculture and more). In order to show this at least partially, the return to an even older article of Phillips than the one from which the discussion of PCs started is of great help.

Phillips (1954) investigated three possible types of fiscal policies, proportional, derivative and integral feedback policy rules (or controls) which change for example government expenditures, broadly speaking, in proportion to output gaps, in proportion to their time rate of change and in proportion to the accumulated differences of such gaps, of course with a negative feedback sign in order to counteract less than normal situations in particular. Similarly, inflation rates may be driven by factor utilization gaps or in the case of wage inflation specifically by deviations of the rate of employment from its NAIRU level, by the rate of change of the employment rate or also by the accumulated differences (where positive and negative signs may occur) of the deviation of unemployment rates from normal levels. Though not framed in this type of language, all three possibilities are in fact taken into account in early or recent investigations of the PC approach, as we shall see in more detail below, the proportional control by the original approach of Phillips, the derivative control in form of the so-called Phillips loops discussion and the paper by Kuh (1967) where the level of wages (and not its rate of change) was related to the rate of unemployment, and the integral control when it was claimed that the rate of unemployment is not in fact determining the rate of inflation itself, but rather its time rate of change. Marrying Phillips (1954) with Phillips (1958) with respect to a treatment of wage and price inflation thus provides a fairly general framework on the basis of which the various findings in the literature on ‘the’ Phillips curve can be evaluated and investigated in a unified way.

Let us first extend our formulation of the wage and price PCs (6.11), (6.12) of Section 6.2 by these additional measures of demand pressure and their influence on wage and price inflation, leaving aside here the same issue for the cost-pressure terms. The wage and price PCs then read:

$$\hat{w} = \beta_{w_1}(V-\bar{V})+\beta_{w_2}\hat{V}+\beta_{w_3}\int_o (V-\bar{V})dt+\kappa_w\hat{p}+(1-\kappa_w)\pi^e \qquad (6.30)$$

$$\hat{p} = \beta_{p_1}(U_c-\bar{U}_c)+\beta_{p_2}\hat{U}_c+\beta_{p_3}\int_o (U_c-\bar{U}_c)dt+\kappa_p\hat{w}+(1-\kappa_p)\pi^e \qquad (6.31)$$

Note here that dimensional homogeneity demands that we should express derivative control in terms of growth rates and not as time rates of change and that it may be preferable to use $V/\bar{V}-1, U_c/\bar{U}_c-1$ as measures of demand pressure (in the place of the simple differences shown above) which at present however only leads to proportional changes in the sizes of the parameters employed in these Phillips curves.

Next there exists another important extension of the considered PCs which takes note of the fact the labor and goods market disequilibrium is in fact reflected in at least two qualitatively different measures of such disequilibrium. In the case of the labor market this is through the external rate of employment of the labor force and the internal rate of employment of the employed, and in the case of the goods market through the rate of utilization of the capital stock and the rate of utilization of the stock on inventories. This leads us to the following alternative extension of the two PCs approach:

$$\hat{w} = \beta_{w_1}(V-\bar{V})+\beta_{w_2}(V^w-1)+\kappa_w\hat{p}+(1-\kappa_w)\pi^e \qquad (6.32)$$

$$\hat{p} = \beta_{p_1}(U_c-\bar{U}_c)+\beta_{p_2}(U_n-1)+\kappa_p\hat{w}+(1-\kappa_p)\pi^e \qquad (6.33)$$

Here $V=L^w/L, V^w=L^d/L^w$ denotes the external and the internal rate of employment and $U_c=Y/Y^p, U_n=N/N^d$ the rate of capacity utilization and the inventory / desired inventory ratio ($L^d$ actual employment in ‘hours’, $L^w$ the employed part of the workforce, also representing the normal working day, and $Y, Y^p$ actual and potential output, $N, N^d$ actual and desired inventory levels). Note here that normal working hours may be diminished by some average rate of ‘absenteeism’ which would imply that the rate of employment of the employed is to be compared with a number smaller than one (in the place of the one used above).

Note also that equations (6.30), (6.31) as well as equations (6.32), (6.33) are of the general form

$$\begin{aligned}\hat{w} &= \beta_w(\cdot) + \kappa_w \hat{p} + (1-\kappa_w)\pi^e \\ \hat{p} &= \beta_p(\cdot) + \kappa_p \hat{w} + (1-\kappa_p)\pi^e\end{aligned}$$

and thus represent (when appropriately reordered) two linear equations in the unknowns $\hat{w}-\pi^e$, $\hat{p}-\pi^e$ that can be uniquely solved for $\hat{w}-\pi^e$, $\hat{p}-\pi^e$ when $\kappa_w, \kappa_p \in [0,1]$ fulfill $\kappa_w\kappa_p < 1$, giving rise then to:

$$\begin{aligned}\hat{w}-\pi^e &= \frac{1}{1-\kappa_w\kappa_p}[\beta_w(\cdot) + \kappa_w\beta_p(\cdot)] \\ \hat{p}-\pi^e &= \frac{1}{1-\kappa_w\kappa_p}[\beta_p(\cdot) + \kappa_p\beta_w(\cdot)].\end{aligned}$$

Integrating across markets for example the two PCs approach (6.32), (6.33) in this way, would thus imply that four qualitatively different measures for demand pressure in the markets for labor as well as for goods have to be used both for money wage and price level inflation for describing their deviation from expected inflation in the usual way by an expectations augmented PC, see Laxton et al. (1998) for a typical example, where, as is customary, only one measure of demand pressure (on the labor market) is considered. Making furthermore use of Phillips’ (1954) three types of control, the obtained integrated PCs will be further differentiated, leading to twelve types of expressions for demand pressure that may appear in the integrated (across markets) price level PC that rules the roost in the mainstream literature. Furthermore, as in Section 6.2, two different types of NAIRUs will then be present in the integrated (wage and) price PC which in general cannot be identified with each other. Finally, as already mentioned, further differentiation may concern the cost pressure terms of the PCs shown above, but will not be considered here in its details.

The stage of wage and price Phillips curves considerations now reached thus exhibits in each case six different measures of demand pressure in the corresponding PC, which, when transformed into integrated PCs, spanning across markets, in the way just shown, leads us to the following fairly complex expressions for expectations augmented PCs:[18]

$$\begin{aligned}\hat{w} = \pi^e &+ \frac{1}{1-\kappa_w\kappa_p}[\beta_{w_1}(V-\bar{V}) + \beta_{w_2}\hat{V} + \beta_{w_3}\int_0 (V-\bar{V})\mathrm{d}t \\ &+ \beta_{w_4}(V^w-1) + \beta_{w_5}\hat{V}^w + \beta_{w_6}\int_0 (V^w-1)\mathrm{d}t\end{aligned}$$

$$
\begin{aligned}
&+ && \kappa_w(\beta_{p_1}(U_c-\bar{U}_c)+\beta_{p_2}\hat{U}_c+\beta_{p_3}\int_0(U_c-\bar{U}_c)\mathrm{d}t \\
&+ && \beta_{p_4}(U_n-1)+\beta_{p_5}\hat{U}_n+\beta_{p_6}\int_0(U_n-1)\mathrm{d}t)] \qquad (6.34) \\
\hat{p}=\pi^e &+ && \frac{1}{1-\kappa_w\kappa_p}[\kappa_p(\beta_{w_1}(V-\bar{V})+\beta_{w_2}\hat{V}+\beta_{w_3}\int_0(V-\bar{V})\mathrm{d}t \\
&+ && \beta_{w_4}(V^w-1)+\beta_{w_5}\hat{V}^w+\beta_{w_6}\int_0(V^w-1)\mathrm{d}t) \\
&+ && \beta_{p_1}(U_c-\bar{U}_c)+\beta_{p_2}\hat{U}_c+\beta_{p_3}\int_0(U_c-\bar{U}_c)\mathrm{d}t \\
&+ && \beta_{p_4}(U_n-1)+\beta_{p_5}\hat{U}_n+\beta_{p_6}\int_0(U_n-1)\mathrm{d}t] \qquad (6.35)
\end{aligned}
$$

As should be obvious now, the second of these equations represents 'the' integrated price Phillips curve of this extended approach to wage and price inflation and its various measures of demand pressure (where the actual wage and price inflation cost-push cross reference has been removed by mathematical substitution). Obviously, this equation is much more complicated than the simple expectations augmented price Phillips curve of the theoretical literature (or its Walrasian reinterpretation as a Lucas supply curve).

Let us briefly consider various applied approaches to PC measurements on the basis of the equations (6.34), (6.35). Fair (1997a,b), as already shown, provides one of the rare studies (disregarding structural macroeconometric model building for the moment) which start from two PCs, though he makes use of $\beta_{p_1} \neq 0$ solely as far as demand pressure variables are concerned. In his view the price Phillips curve is therefore the important one.

Concerning, modern macroeconometric model building, we find in Powell and Murphy (1997) a money wage Phillips curve with $\beta_{w_1}, \beta_{w_2} \neq 0$ and a price Phillips curve that appears to be based on cost-push terms solely, but which (when appropriately reformulated, see Chiarella, Flaschel, Groh, Köper and Semmler (2000a)) in fact also makes use of $\beta_{p_1} \neq 1$ implicitly. Furthermore, the parameter $\beta_{w_2}$ is in their study about eight times larger than $\beta_{w_1}$ when the nonlinear wage Phillips curve measured in this work is linearized at the steady state, which supports Kuh's (1967) assertion that the wage Phillips curve is a level relationship rather than one concerning rates of inflation (and which at the same time stresses the importance of Phillips loops as already observed by Phillips (1958) himself). Indeed, if $\hat{w} = \beta_{w_2}\hat{V}$ represents the dominant part of the money wage Phillips curve, we get by integration $w = constV^{\beta_{w_2}}$ and thus a wage curve as considered on the microlevel by Blanchflower and Oswald (1994) for example. In this view the wage Phillips curve, with derivative control solely, is therefore the important one.

Laxton et al. (1998) use for the Multimod mark III model of the IMF an integrated (or hybrid) PC of the type (6.35) with only $\beta_{w_1} \neq 0$, and thus the most basic type of PC approach, but stress instead the strict convexity of this curve and the dynamic NAIRU considerations this may give rise to. In their view, therefore, the wage Phillips curve with proportional term only is the important one. Stock and Watson (1997) find evidence for a Phillips curve of the type $\dot{\pi} = \beta_{w_3}(V - \bar{V})$, $\pi = \hat{p}$, which shows that this view is in fact based on an integral control in the money wage Phillips curve (solely) and possibly also on an implicit treatment of inflationary expectations in addition. Roberts (1997) derives a conventional expectations augmented price Phillips curve from regional wage curves as in Blanchflower and Oswald (1994) and thus argues that proportional control is relevant in the aggregate even if derivative control applies to the regional level.

We thus find in this brief discussion of applied approaches a fairly varied set of opinions, which is, however, not so varied as to pay attention to inside employment rates and inventory utilization rates and which only in the case of Fair (1997a,b) takes account of the possibility that demand pressure on the goods market may be qualitatively and quantitatively different from demand pressure on the labor market with respect to extent and implications. Otherwise, however, at least the possibility for proportional, derivative and integral control is taken into account by this literature (though not reflected and compared in these terms). It must therefore be noted that the discussion on Phillips curves is at present again a lively one, a still unsettled one, but also one with still a very limited horizon. Of course, not all of the expressions shown in (6.35) must be relevant from the empirical point of view, at all times and in all locations. But this should be the outcome of a systematic investigation and not the result of more or less isolated views and investigations of already very specialized types of PCs. Despite the new approaches to PC analysis it therefore appears as if the analysis and investigation of these curves should start anew from the extended perspective we have tried to describe above.

Let us close this section by considering a theoretical approach by Rowthorn (1980) which makes use of a price Phillips curve with proportional control and a wage Phillips curve with derivative control in order to provide an IS–LM–PC model, in his case in fact a monetarist model of inflation and stagflation, which is formally of the same type as the ones considered in the preceding section, but which allows for an endogenous determination of the NAIRU rates $\bar{V}, \bar{U}_c$ (based on the conflict over income distribution). This is an interesting extension of the IS–LM–PC dynamics considered in Section 6.3 and it furthermore provides a theoretical example on how the use of various special types of Phillips curves (appropriately combined) can lead to quite different views of the interaction of unemployment and inflation as compared to the conventional one.[19]

The fundamental features and building blocks of Rowthorn's reformulation of this interaction (here augmented by IS–LM analysis in the place of his simpler quantity theoretic approach) are the following ones:

$$\hat{p} = \beta_p(\Pi^* - \Pi) + \pi^e \tag{6.36}$$

$$\Pi^* = \Pi^*(U_c) \tag{6.37}$$

$$\Pi = 1 - u, \quad u = (w/p)l_y \quad \text{the share of wages} \tag{6.38}$$

$$\hat{w} = \beta_w \hat{V} + \hat{p} \tag{6.39}$$

We have a price Phillips curve of the proportional kind (based on a kind of self-reference to price inflation expected to hold over the medium run) and a wage Phillips curve of the derivative type with myopic perfect foresight as far as price inflation is concerned. Price inflation is driven by the gap between the desired profit share $\Pi^*$ and the actual one, $\Pi$, with the desired profit share being a positive function of the rate of capacity utilization $U_c$ of firms. In the background of this model we have our fixed proportions technology with given labor productivity $1/l_y$ and thus get from this a strict proportionality between the rate of capacity utilization and the rate of employment, as a very simple form of Okun's law, $U_c = const \cdot V$, as shown in Section 6.3. Furthermore, the money wage Phillips curve gives rise to (by its integration): $w/p = const\ V^{\beta_w}$, a functional form that then also applies to the wage share $u$ in the place of the real wage $w/p$. Inserting all these expressions into the price Phillips curve $\hat{p} = \beta_p(\Pi^* - \Pi) + \pi^e$ gives rise to

$$\hat{p} = \beta_p(\Pi^*(V) - (1 - u(V))) + \pi^e = \beta_p(\Pi^*(V) + u(V) - 1) + \pi^e$$

with both $\Pi^*, u$ being strictly increasing functions of the rate of employment $V$. On the surface this is just an ordinary PC of the monetarist type (as we have employed in Section 6.3), though now possibly a nonlinear one. Disregarding this latter possibility and assuming that parameters are such that there is a solution $\bar{V} \in (0, 1)$ where $\Pi^*(\bar{V}) + u(\bar{V}) = 1$ holds (which is then uniquely determined), we then get from these alternative underpinnings of the IS–LM–PC model analyzed in Section 6.3 an endogenous explanation of the NAIRU rate of employment, which was there given as a parameter. This NAIRU rate, among others, now depends on the relationship $\Pi^*(U_c)$, and thus negatively on the steepness of this curve (which characterizes the strength with which capital owners defend their income shares) and also negatively on the parameter $\beta_w$ which measures the strength with which labor defends its income share. Therefore, the stronger the conflict over income distribution, the lower is the NAIRU rate of employment at which the income shares demanded, $\Pi^*(V) + u(V)$, become compatible with what is available for distribution, thereby allowing for a steady behavior of wage and price inflation. This provides in simple terms a simultaneous interpretation of both the NAIRU rate of capacity utilization and the NAIRU rate of employment, surrounded by dynamics that are of the same type as the one investigated in the preceding section. We stress again that this has become possible through a simple specialization of the very general type of PCs we have investigated in the present section. Note however that Okun's law, which is based on a positive correlation between the rate of capacity utilization and the rate of employment, has been used here in order to derive this specific view on the explanation of steady state rates of factor utilization.

We conclude this section with the observation that much remains to be done in the theoretical discussion of the form and the implications of PC approaches, where many more outcomes may be obtained than is generally believed. The same holds true for empirical studies of Phillips curves, where there is a lack of systematic investigation of the wealth of possibilities our extended presentation of the wage price module of integrated dynamical models of Keynesian or other variety can give rise to. Furthermore, what has been discussed here for PCs can also be applied to the

ACs considered in Section 6.2, see Figure 6.3, where derivative control terms for the impact of capacity utilization on capital stock and employment growth would introduce Harrodian accelerator aspects into the growth cycle there considered and where integral terms (for profitability) would represent one possibility to introduce medium run aspects into such AC analysis.

## 6.5 The basic integrated KMG model of fluctuating growth

Let us now put together what we have learned about AC-PC analysis and IS–LM–PC analysis with their common element, the wage price adjustment equations. We shall use these wage price equations in the form (6.11), (6.12) used in Section 6.2 and will thus considerably generalize the conventional representation of the expectations augmented integrated price Phillips curve of Section 6.3, but will not yet take the generalizations of this approach proposed in Section 6.4 into account here. Instead, our presentation of an integrated IS–LM–PC-AC or K(eynes)–M(etzler)–G(oodwin) model will add to Section 6.2 an investment function that now includes, besides a measure of the profitability of firms, the real rate of interest as in IS–LM–PC analysis and the rate of capacity utilization as in Malinvaud (1980), but not yet in derivative form as in Harrodian knife edge growth analysis. Furthermore, we will have two types of households in the following, workers who consume what they get and asset holders who consume and save (the latter in the form of money, government bonds and equities). Finally, we will reformulate and extend this synthesis of AC-PC and IS–LM–PC analysis by means of a Metzlerian inventory adjustment mechanism and thus allow, in addition to sluggish wage and price adjustment, also for sluggish quantity adjustment. In the place of IS-equilibrium we thus will have the interaction of sales expectations, output (including desired inventory changes) and actual aggregate demand and on this basis unintended inventory changes and an inventory adjustment mechanism. Since this model wants to be complete with respect to the interaction of all these elements we have, of course, to consider the capacity effects of investment in addition to the income effects of investment and thus to include growth which we here supplement by natural growth of labor force in the usual way (and implicitly also by Harrod neutral technical change, which does not alter the presentation of the model, but demands only that all labor and wage expression have to reinterpreted in terms of efficiency units).

The model we obtain in this way is the working model of Chiarella and Flaschel (2000b) which is derived in their book as a systematic extension of the earlier traditional models of monetary growth of Tobin, Keynes–Wicksell and IS–LM type. This model type is easily extended to include, besides the neutral form of technical change just mentioned, a more refined government sector, in particular with respect to income taxation, smooth factor substitution in the place of the fixed proportions technology of this chapter, more advanced modeling of expectation formation than are considered in this section as a generalization of Section 6.3's adaptive expectations mechanism, and more. These generalization are considered in their relevance and with respect to their implications in Chiarella and Flaschel (2000a) amd Chiarella, Flaschel, Franke and Lux (2001) and will here be left aside for reasons of simplicity, since they do not alter the conclusions drawn in this section in a significant way.

The working model of Chiarella and Flaschel (2000b) that we reconsider below, here called for simplicity the KMG model due to its synthesis of IS–LM analysis (Keynes, 1936, Hicks, 1937), delayed goods market adjustment processes (Metzler, 1941) and the classical growth cycle mechanism (Goodwin, 1967), represents in our view the basic format for traditional Keynesian monetary growth analysis with which all extensions, modifications, reformulations or completely new formulations of monetary growth models not based on Say's Law should be compared and evaluated. We are of course aware that the Hicksian representation of the Keynes-component of this model type will not be considered as a proper Keynes-representation by a variety of macroeconomists. Nevertheless it is important to have such a traditional reference case, the KMG model, at our disposal in order to allow for a precise presentation of where one should depart from it in order to get a better theory of fluctuating growth of deterministic or even of stochastic type. The problem in the literature on monetary growth, see Orphanides and Solow (1990) for example, was, that such a traditional integrated Keynesian prototype dynamics was completely missing as an explicitly spelled out model type, not to speak of the analysis of the six dimensional dynamics this model type will give rise to (which in fact is the minimum dimension of a theory of monetary growth with sluggish adjustments in wages, prices, expectations, quantities and inventories).

We start with an overview on the intensive form representation of these 6D KMG-dynamics. The extensive form of this KMG model is presented in its details in Chiarella and Flaschel (2000b), where it is shown that the intensive form results from a model type that is coherently formulated with respect to behavior and budget restrictions and thus does not allow for demand of agents that is not backed up by the supply of funds (or loans) that finance these expenditures.

The first dynamical law concerns the real wage $\omega = w/p$ and is in fact a direct consequence of the integrated two PCs approach (6.7), (6.8), as it has been derived from the wage–price dynamics (6.11), (6.12), now including inflationary expectations $\pi^e \neq 0$ which however cancel in the calculation of the real wage. Subtracting (6.8) from (6.7) gives immediate rise to

$$\begin{aligned}\hat{\omega} &= \hat{w} - \hat{p} \\ &= \frac{1}{1-\kappa_w\kappa_p}[(1-\kappa_p)\beta_w(V-\bar{V}) - (1-\kappa_w)\beta_p(U_c-\bar{U}_c)] \qquad (6.40)\end{aligned}$$

which simply states that the adjustment of real wages depends positively on the demand pressure on the market for labor and negatively on that in the market for goods. We immediately realize that either wage or price flexibility should bring instability to the dynamics of this section (normal or abnormal Rose effects), depending on whether real wage increases increase or decrease economic activity.

The next law of motion concerns the dynamics of the factor endowment ratio $l = L/K$, where it is assumed that labor supply $L$ grows with the given natural rate

$n$ and the capital stock (in our Keynesian approach) with the rate of net investment $\hat{K} = I/K$. The essential element in this law of motion is therefore given by the investment function which is specified as follows:

$$I/K = i_1(\rho^e - (r - \pi^e)) + i_2(U_c - \bar{U}_c) + n.$$

This function (as all other behavioral equations) is assumed as linear (just as our simple production function $L^d = l_y Y, Y^p = y^p K, l_y, y^p$ is given magnitudes) in order to keep the model first as linear as is possible which allows us to concentrate on its intrinsic or unavoidable nonlinearities in the beginning of the analysis of KMG growth dynamics. Note also that the trend term in this investment equation is given by the natural rate of growth.

Investment per unit of capital thus depends on the expected rate of profit $\rho^e$, the real rate of interest $r - \pi^e$ and the rate of capacity utilization $U_c$, representing Tobin's $q$ and the capacity effect considered by Malinvaud (1980) and others. All these magnitudes are actual (short run) values and should be replaced by medium run averages in applications of this model type, which increases the dimension of the dynamics without adding too much new structure to the model, see Flaschel, Gong, and Semmler (1999) for such extensions and applications of the model. Here, however, only the above simple formulation of investment behavior will be used and it gives rise to the following law of motion for the labor intensity $l = L/K$ :

$$\widehat{l} = -i_1(\rho^e - (r - \pi^e)) - i_2(U_c - \bar{U}_c). \tag{6.41}$$

The next dynamical law is basically equation (6.12) of Section 6.2, since it is based on the definitional equation for real balances (per unit of capital now): $m = M/(pK)$, implying $\hat{m} = \bar{\mu} + n - \hat{l} - \hat{p}$, where it is again assumed that the money supply $M$ grows with a given rate $\bar{\mu}$ (since policy questions are not of interest here). Making use of the equation (6.12) for $\hat{p}$ this gives:

$$\widehat{m} \quad = \quad \bar{\mu} - n + \hat{l} - \pi^e - \frac{1}{1 - \kappa_w \kappa_p}[\beta_p(U_c - \bar{U}_c) + \kappa_p \beta_w(V - \bar{V})] \tag{6.42}$$

We see that increased capacity utilization on both the labor and the goods market will speed up inflation and thus reduce the growth rate of real balances, leading to corresponding nominal interest rate changes due to the Keynes-effect as it derives from the simple LM-curve still present in this model type.

Next we have the law for inflationary expectations which is a simple extension of the one used in Section 6.3, since we now determine these expectations as an average of backward and forward looking behavior (time series methods and forecasts by means of small theoretical models). Time series methods can in principle be as complicated and refined as possible, when only numerical simulations of the model are intended. From the viewpoint of theory they should at first be chosen to be as simple as possible in order to allow for an analytical treatment of stability issues as we shall provide below, i.e. they will be of the simple adaptive type made use of already in Section 6.3 (they can be made a humped shaped average of past observations of inflation by means of nested adaptive expectation schemes in a next step for example). Forward looking expectations can be based on the p-star concept of the FED and the German Bundesbank for example, which says that inflation rates will converge to the difference of $\bar{\mu}$, the growth rate of the money supply, and $\hat{Y}^p$, the growth rate of potential output (as long as the velocity of money can be considered a given magnitude). Made as simple as possible again this shows that inflationary expectations of this type assume that there is convergence of these actual inflation rates back to the steady rate $\bar{\mu} - n$ giving rise to:

$$\begin{aligned} \dot{\pi}^e &= \beta_\pi[\alpha(\hat{p} - \pi^e) + (1-\alpha)(\bar{\mu} - n - \pi^e)] \qquad (6.43) \\ &= \beta_\pi[\alpha \frac{1}{1-\kappa_w\kappa_p}[\beta_p(U_c - \bar{U}_c) + \kappa_p\beta_w(V - \bar{V})] \\ &+ (1-\alpha)(\bar{\mu} - n - \pi^e)] \end{aligned}$$

where $\alpha$ denotes the weight attached to the backward looking type of expectations and $1-\alpha$ the one for the forward looking type. The destabilizing role of the Mundell effect is clearly visible in this extended equation, since economic activity depends positively on expected inflation (due to the assumed investment behavior) and since an increase in economic activity, here measured by two rates of factor capacity utilization, speeds up the increase in inflationary expectations as shown by this equation.

There remains the quantity adjustment process on the market for goods which is driven by the adjustment of sales expectations and the changes in actual inventories, $y^e = Y^e/K$ and $\nu = N/K$, both already in per unit of capital form here, see Metzler (1941) for the original approach. The two laws of motion for these variables read:

$$\begin{aligned} \dot{y}^e &= \beta_{y^e}(y^d - y^e) + \hat{l}y^e \qquad (6.44) \\ \dot{\nu} &= y - y^d + (\hat{l} - n)\nu \qquad (6.45) \end{aligned}$$

where the terms involving $l$ are simply due to the fact that everything is expressed in per unit of capital terms. Sales expectations of firms, $y^e$, are here assumed to change in an adaptive fashion, following actual demand $y^d$ with some time delay, while actual inventories changes are given by definition through the difference between actual output $y$ and actual demand $y^d$, again corrected by a term that takes account of the intensive form under consideration.

This closes the description of the laws of motion of the state variables of our basic KMG dynamics which concern income distribution, relative factor growth, inflation as measured by the change of real balances, inflationary expectations, sales expectations and actual inventory changes. These dynamical laws do not yet form a complete system, but must be supplemented by some algebraic equations which define the statically endogenous magnitudes we used in the above differential equations. They are given by:

$$\begin{aligned} y &= \beta_n(\beta_{n^d}y^e - \nu) + (1 + n\beta_{n^d})y^e \\ y^d &= \omega l_y y + (1 - s_c)(\rho^e - t^n) + i_1(\rho^e - (r - \pi^e)) + i_2(U_c - \bar{U}_c) + n + \delta + g \end{aligned}$$

$$
\begin{aligned}
V &= l^d/l, \; U_c = y/y^p, \; l^d = L^d/K = l_y y, \\
\rho^e &= y^e - \delta - \omega l_y y, \; r = r_0 + (h_1 y - m)/h_2.
\end{aligned}
$$

These equations describe output $y$ equal to expected sales and voluntary inventory changes, which follow $\beta_{n^d} y^e - \nu$, the difference between desired inventories and the actual ones, with speed $\beta_n$, aggregate demand $y^d$ which is composed of the real wage sum per unit of capital, consumption of asset holders based on expected profits after taxes $\rho^e - t^n$, where taxes net of interest $t^n$ are treated as a parameter of the model as in Sargent (1987, Ch.5), gross investment ($\delta$ the depreciation rate) and government expenditures $g$, again assumed a parameter of the model. The remaining equations then define the rate of employment $V$, the rate of capacity utilization $U_c$, the expected rate of profit, $\rho^e$, based on the sales expectations of firms, and the nominal rate of interest as defined by a linear form of money market equilibrium. These explanations must suffice here as a presentation of the static part of the working KMG model of Chiarella and Flaschel (2000b). For additional explanations and a detailed analysis of this model type the reader is referred to this earlier work.

Note with respect to the full structure of traditional Keynesian monetary growth analysis presented in Section 6.1, that we now employ an advanced description of the wage–price module (but not as advanced as the one arrived at in Section 6.4), that we still consider a closed economy (no Dornbusch exchange rate dynamics), that internal asset markets (up to a static LM determination of the nominal rate of interest) are not considered explicitly (as in Sargent (1987, Ch.1–5), i.e. we do not yet consider Blanchard (1981) type asset market dynamics, we do not yet have Fisher debt effects, as considered in Section 6.2 and have also abstracted from wealth (Pigou) effects in consumption as well as money demand. The above disequilibrium growth model therefore just represents the beginning of the analysis of integrated disequilibrium growth of a traditional Keynesian type, with particular stress on sluggish wage, price and quantity dynamics and resulting capacity utilization problems for labor as well as capital, but not yet a real-financial interaction in the proper sense of this phrase.

Despite the linearity assumed for all behavioral equations the model is intrinsically nonlinear, due to its use of growth laws of motion and the unavoidable appearance of products and quotients of state variables in various places. The considered dynamics therefore should be capable of generating limit cycles and also more complex attractors for its trajectories even on this basic level of its formulation.

**Proposition**

1. *Assume sufficiently sluggish adjustment for wages, prices, and inflationary expectations, and a strong Keynes-Effect ($h_2$ small). Then:*
   *The interior steady state of the 6D dynamics (6.40) – (6.45), which is easily calculated and uniquely determined, is locally asymptotically stable for all adjustment speeds of sales expectations $\beta_{y^e}$ chosen sufficiently large and speeds of adjustments of inventories, $\beta_n$, chosen sufficiently low.*

2. *The 6D determinant of the Jacobian of the dynamics at the steady state is always positive.*

3. *If $\beta_{\pi^e}, \beta_n, h_2$ are chosen sufficiently large, the steady state equilibrium becomes locally repelling. The system therefore undergoes a (generally non degenerate) Hopf bifurcation at intermediate value of these (and other) parameters, which generates persistent fluctuations, that are attractors in the supercritical case and repellers in the subcritical case.*

The periodic fluctuations obtained in this way integrate the growth cycle analysis of Section 6.2 with the inflationary dynamics of Section 6.3, coupled with Metzlerian quantity adjustment in the market for goods and they are generated independently of any kink in the money wage PC. Further details on this proposition are provided in Chiarella and Flaschel (2000b). We do not go into a proof of this proposition here, but simply add some explanations to the assertions made. The steady state of the system is asserted to be locally attracting for all price and quantity adjustment speeds (including the Metzlerian inventory accelerator) sufficiently low (up to sales expectations which mirror the stable Keynesian multiplier dynamics which improve stability if chosen sufficiently large). In addition we should have a fairly interest-inelastic money demand function in order to produce large positively correlated swings of the nominal rate of interest when the price level rises or falls. Partial insights on the stability of Keynesian dynamics (augmented with what we know for the Rose effects considered in Section 6.2) or even static conclusions of Keynesian theory, appropriately combined, thus allow here for a stability assertion for the full 6D dynamics of the integrated KMG growth model. Furthermore, since the determinant does not change sign when the parameters of the model are changed, we know that loss of stability can only occur in a cyclical fashion since eigenvalues must then cross the imaginary axis excluding 0 (and will generally do so with positive speed). The resulting situations of Hopf bifurcations then generally imply that this change in stability of the system is accompanied by either the ‘birth’ of a stable limit cycle (with increasing amplitude) to the right of the critical bifurcation value (where a pair of eigenvalues has become purely imaginary) or the death of an unstable limit cycle (via its shrinking amplitude and the disappearance of a ‘stable corridor’) to the left of this critical point when this point is approached.

The above brief considerations of KMG growth dynamics must here suffice to indicate what results might be expected from an integrated AC-PC analysis with Keynesian supply rationing and with also sluggish quantity adjustment processes of Metzlerian type. These topics are further pursued in particular in Chiarella and Flaschel (2000b) and Chiarella, Flaschel, Groh and Semmler (2000).

## 6.6 Conclusions and outlook

We have started in this chapter from a simple classical growth cycle model, based on AC-PC analysis, and from the textbook understanding of IS–LM analysis augmented by inflation and expected inflation in the conventional way, here called

IS–LM–PC analysis. We could show for this latter model type that textbook presentations of it are generally misleading and incomplete and give rise, when corrected and completed, to persistent fluctuations of employment and inflation around the steady state position when adjustment of adaptively formed expectations is sufficiently fast. These two model types therefore provide us with a theory of medium run as well as long run employment cycles which, due to their common PC element suggests, that it should be possible to integrate them into larger dynamics with at least two cycle generating mechanisms. We have supplemented this discussion of growth and inflation with a reconsideration of theoretical and applied PC discussions and have found, that the former is still fairly underdeveloped with respect to a general treatment of wage–price dynamics and the demand pressure elements that may be involved in this interaction, that applied approaches have generally been broader here, but did not treat their empirical investigations from the unified point of view that we could present in Section 6.4.

In Section 6.5 we then presented a coherent core model of short run Keynesian quantity dynamics, medium run wage–price dynamics and long run growth cycle dynamics, based on fully specified budget equations for all sectors considered, see Chiarella and Flaschel (2000b), and on a unique economic steady state reference path, which integrated the Goodwin (1967) AC-PC analysis and the Rose (1967) real wage effects considered in Section 6.2 with the IS–LM–PC analysis (based on Keynes- and Mundell-effects) of Section 6.3, by making use in addition of Metzlerian quantity adjustment processes on the market for goods in the place of an infinitely fast adjustment to Keynesian goods market equilibrium. This model type, the working model of Chiarella and Flaschel (2000b), there obtained from a hierarchical structured sequence of disequilibrium models of monetary growth, should long since have been the basic reference case for an integrated treatment of Keynesian disequilibrium growth theory, but has indeed been completely neglected in the literature on monetary growth, see Orphanides and Solow (1990) for example, possible due to the fact that AS–AD growth, as treated in Sargent (1987, Ch. 5) was generally considered as the Keynesian theory of monetary growth and due to the fact that the analysis of nonlinear economic dynamics seemed to be restricted to dimension 3 or less, while our integrated model of Keynesian monetary growth, with sluggish price and quantity adjustment (including expectation formation on aggregate demand as well as inflation), and under- or over-utilized labor as well as capital is in the minimum of dimension 6.

The scene is thereby set for a fundamental and at the same time already very general labor and goods market disequilibrium growth model (with a simple LM treatment of asset markets still) which can be analyzed from the theoretical as well as from the numerical and empirical perspective, see Chiarella and Flaschel (2000b) and Flaschel, Gong and Semmler in particular, and which should serve at the least as a reference case for judging the progress made by other types of approaches in the recent past, by those who believe that Keynesian monetary growth theory (which was never fully presented and analyzed as we have shown) is nowadays dated or even outdated. In our view such a judgment is not based on knowledge of the true potential of integrated Keynesian theory of traditional type, as we have attempted to show here and in more detail in Chiarella and Flaschel (2000b). Chapters 3 to 5 in Chiarella, Flaschel, Groh and Semmler (2001) show in this respect furthermore, that recent non-Walrasian macro-approaches with their regime switching methodology can easily be used to make the traditional KMG approach even more coherent, by adding to it supply side constraints of a fairly secondary nature, and that also the partial new Keynesian treatment of market imperfections can be integrated into this KMG approach or used to overhaul certain of the modules of this approach from these partial perspectives.

There are further topics which are relevant for the discussion presented in this chapter but which are not considered here due to space limitations. Opening the KMG economy is the subject of Chiarella and Flaschel (1999, 2000c) and Chiarella, Flaschel, Franke and Lux (2001), developing the model further towards applied integrated macrodynamics is the subject of a series of papers, Chiarella and Flaschel (2000d–f), Chiarella, Flaschel and Zhu (2000a-c), Chiarella, Flaschel, Groh, Köper and Semmler (2000a,b) and debt deflation is further analyzed in a very general model type in Chiarella and Flaschel (2000d,g ), Chiarella, Flaschel and Semmler (2000, 2001). Taking this work together I hope will lead to the formulation of an applicable theoretical model of integrated disequilibrium growth with all the features shown in Figure 6.1 and thus of a fully specified traditional Keynesian model type that is consistently formulated with respect to budget equations and steady state calculations and that allows for a detailed analysis of the various feedback hypotheses that can be associated with a structure as shown in Figure 6.1. This figure thus represents the perspective of this chapter which I believe has shown that such an aim can be pursued by a systematic integration and extension of classical AC-PC and Keynesian IS–LM–PC analysis.

## Notes

1 See Chiarella and Flaschel (2000h) for the full details of what is sketched in this chapter and in the concluding section in particular.
2 continuity of this curve is of course assumed in addition.
3 Note that we reinterpret the NAIRU of the literature here in terms of the rate of employment (or utilization) $V$ of the labor force, not in terms of unemployment.
4 Note that a real wage PC is obtained from the conventional money wage PC (augmented by inflationary expectations of course) by assuming myopic perfect foresight with regard to the expected rate of inflation.
5 See Flaschel, Franke and Semmler (1997) for the consideration of cross-dual macrodynamics on various levels of generality.
6 See also Flaschel (1984) in this regard.
7 We assume that the parameters of the model are such that both $u_o$ and $d_o$ are positive.
8 See Chiarella, Flaschel amd Semmler (2000) for a much more general treatment of such growth dynamics.
9 We assume again that parameters are chosen such that all steady state values are meaningful.
10 These authors employ a special discrete time version of the following model, a difference which however is not essential for our following discussion of their model.
11 Note that the real wage is constant in this approach which is therefore clearly complementary to one in the preceding section.

12 which in principle should be well-known from Tobin (1975) and subsequent work, but which is still unfamiliar, see Groth (1993).

13 Note that – though globally asymptotically stable – the model is still incomplete since the right half of the phase plane is not an invariant set of this dynamics, i.e. output can be become negative along trajectories that start in an economically meaningful domain.

14 See Laxton, Rose and Tambakis (1997) for an empirical discussion of the kind of nonlinearity that may be involved in the integrated price level – rate of unemployment Phillips curve.

15 The two systems are identical at the border line $\hat{w} = \beta_w(V - \bar{V}) + \pi^e = 0$.

16 See here also Flaschel (1994) and Flaschel and Franke (1996,98).

17 See for example Barnett and He (1998) for an exception.

18 Note that $\pi^e$ was set equal to zero in the models considered in Section 6.2.

19 See also Flaschel (1993) and Flaschel and Groh (1996) in this regard.

## References

Barnett, W. and Y. He (1998) 'Bifurcations in continuous-time macroeconomic systems'. Washington University in St. Louis: Mimeo.

Barro, R. (1994) 'The Aggregate-Supply Aggregate-Demand Model'. *Eastern Economic Journal*, 20, 1–6.

Blanchard, O.J. (1981) 'Output, the stock market, and interest rates'. *American Economic Review*, 71, 132–43.

Blanchard, O. (1997) 'Is there a core of usable macroeconomics?' *American Economic Review. Papers and Proceedings*, 87, 244–6.

Blanchflower, D.G. and A.J. Oswald (1994) *The Wage Curve*. Cambridge, MA: The MIT Press.

Chiarella, C. and P. Flaschel (1999) 'Keynesian monetary growth in open economies'. *Annals of Operations Research*, 89, 35–59.

Chiarella, C. and P. Flaschel (2000a) 'The emergence of complex dynamics in a 'naturally' nonlinear integrated Keynesian model of monetary growth'. In: W. A. Barnett, C. Chiarella, S. Keen, R. Marks and H. Schnabl (eds) *Commerce, Complexity, and Evolution: Topics in Economics, Finance, Marketing and Management.* Proceedings of the Twelfth Symposium on Economic Theory and Econometrics. Cambridge: Cambridge University Press.

Chiarella, C. and P. Flaschel (2000b) *The Dynamics of Keynesian Monetary Growth: Macrofoundations*. Cambridge: Cambridge University Press.

Chiarella, C. and P. Flaschel (2000c) "'High order" disequilibrium growth dynamics: Theoretical aspects and numerical features'. *Journal of Economic Dynamics and Control*, 24, 935–63.

Chiarella, C. and P. Flaschel (2000d) 'Towards applied disequilbrium growth theory: I. The starting theoretical model'. Discussion Paper: University of Technology, Sydney.

Chiarella, C. and P. Flaschel (2000e) 'Towards applied disequilbrium growth theory: II. Intensive form and steady state analysis of the model'. Discussion Paper: University of Technology, Sydney.

Chiarella, C. and P. Flaschel (2000f) 'Towards applied disequilbrium growth theory: III. Basic partial feedback structures and stability issues'. Discussion Paper: University of Technology, Sydney.

Chiarella, C. and P. Flaschel (2000g) 'Applying disequilibrium growth theory. II. Housing cycles and debt deflation'. Mimeo: University of Bielefeld.

Chiarella, C. and P. Flaschel (2000h) 'Disequilibrium growth theory: Foundations, synthesis, perspectives'. Discussion paper: University of Techonology, Sydney.

Chiarella, C., P. Flaschel, R. Franke (2001) *Integrated Macroeconomic Dynamics: Foundations, Extensions, Perspectives*. Book manuscript, in preparation.

Chiarella, C., P. Flaschel, R. Franke and T. Lux (2001) *Open Economy Macrodynamics: The Keynes–Metzler–Goodwin Approach*, (forthcoming). Heidelberg: Springer Verlag.

Chiarella, C., P. Flaschel, G. Groh, C. Köper and W. Semmler (2000a) 'Towards applied disequilibrium growth theory: VI. Substitution, money-holdings, wealth-effects and further extensions'. Discussion paper: University of Technology, Sydney.

Chiarella, C., P. Flaschel, G. Groh, C. Köper and W. Semmler (2000b) 'Towards applied disequilibrium growth theory: VII. Intensive form analysis in the case of substitution'. Discussion Paper: University of Technology, Sydney.

Chiarella, C., P. Flaschel, G. Groh, W. Semmler (2000) *Disequilibrium, Growth and Labor Market Dynamics: Macro Perspectives.* Heidelberg: Springer Verlag.

Chiarella, C., P. Flaschel, W. Semmler (2001) 'The macrodynamics of debt deflation'. In: R. Bellofiore and P. Ferri (eds) *Financial Fragility and Investment in the Capitalist Economy* Aldershot: Edward Elgar .

Chiarella, C., P. Flaschel, W. Semmler (2001) 'Price flexibility and debt dynamics in a high order AS–AD model' To appear: *Central European Journal of Operations Research*.

Chiarella, C., P. Flaschel and P. Zhu (2000a) 'Towards applied disequilibrium growth theory: IV. Numerical investigations of the core 18D model'. Discussion Paper: University of Technology, Sydney.

Chiarella, C., P. Flaschel and P. Zhu (2000b) 'Towards applied disequilibrium growth theory: V. Extrinsic nonlinearities and global boundedness'. Discussion Paper: University of Technology, Sydney.

Chiarella, C., P. Flaschel and P. Zhu (2000c) 'Towards applied disequilibrium growth theory: VIII. A 22D core dynamics in the case of substitution'. Discussion Paper: University of Technology, Sydney.

Chiarella, C., P. Flaschel, P. Zhu (2000d) 'Towards applied disequilibrium growth theory: IX. Global boundedness in the case of substitution'. Discussion Paper: University of Technology, Sydney.

Desai, M. (1973) 'Growth cycles and inflation in a model of the class struggle'. *Journal of Economic Theory*, 6, 527–45.

Dornbusch, R. and S. Fischer (1996) *Macroeconomics.* New York: McGraw-Hill (7th edition).

Fair, R. (1997a) 'Testing the NAIRU model for the United States'. Yale University: Mimeo.

Fair, R. (1997b) 'Testing the NAIRU model for 27 countries'. Yale University: Mimeo.

Fisher, I. (1933) 'The debt-deflation theory of great depressions'. *Econometrica*, 1, 337–57.

Flaschel, P. (1984) 'Some stability properties of Goodwin's growth cycle model'. *Zeitschrift für Nationalökonomie*, 44, 63–9.

Flaschel, P. (1993) *Macrodynamics. Income Distribution, Effective Demand and Cyclical Growth*. Frankfurt/M.: Verlag Peter Lang.

Flaschel, P. (1994) 'A Harrodian knife-edge theorem for the wage–price sector?' *Metroeconomica*, 45, 266–78.

Flaschel, P. and R. Franke (1996) 'Wage flexibility and the stability arguments of the Neoclassical Synthesis'. *Metroeconomica*, 47, 1–18.

Flaschel, P. and R. Franke (2000) 'An old-Keynesian note on destabilizing price flexibility'. *Review of Political Economy*, 12, 273–83.

Flaschel, P., Franke, R. and W. Semmler (1997) *Dynamic Macroeconomics: Instability, Fluctuations and Growth in Monetary Economies.* Cambridge, MA: The MIT Press.

Flaschel, P., Gong, G. and W. Semmler (1999) ‘A Keynesian based econometric framework for studying monetary policy rules’. University of Bielefeld: Mimeo.

Flaschel, P. and G. Groh (1996) *Keynesianische Makroökonomik. Unterbeschäftigung, Inflation und Wachstum.* Heidelberg: Springer.

Flaschel, P. and G. Groh (1998) ‘Textbook Stagflation Theory and Beyond’. University of Bielefeld: Discussion Paper.

Goodwin, R.M. (1967) ‘A growth cycle’. In: C.H. Feinstein (ed.): *Socialism, Capitalism and Economic Growth.* Cambridge: Cambridge University Press, 54–8.

Groth, C. (1993) ‘Some unfamiliar dynamics of a familiar macro model’. *Journal of Economics*, 58, 293–305.

Hicks, J. (1937) ‘Mr Keynes and the Classics’, *Econometrica*, 5, 147–59.

Keen, S. (1999) ‘The nonlinear economics of debt deflation’. In: W.A. Barnett, C. Chiarella, S. Keen, R. Marks, and H. Schnabl (eds): *Commerce, Complexity and Evolution: Topics in Economics, Finance, Marketing and Management.* Proceedings of the Twelfth International Symposium in Economic Theory and Econometrics. Cambridge: Cambridge University Press.

Keynes, J.M. (1936) *The General Theory of Employment, Interest and Money.* New York: Macmillan.

Kuh, E. (1967) ‘A productivity theory of the wage levels – An alternative to the Phillips curve’. *Review of Economic Studies*, 34, 333–60.

Laxton, D., D. Rose and D. Tambakis (1997) ‘The U.S. Phillips curve: The case for asymmetry’. Mimeo.

Laxton, D., P. Isard, H. Faruquee, E. Prasad and B. Turtelboom (1998) *MULTIMOD Mark III. The core dynamic and steady-state models.* Washington, DC: International Monetary Fund.

Malinvaud, E. (1980) *Profitability and Unemployment*. Cambridge: Cambridge University Press.

Metzler, L. A. (1941) ‘The nature and stability of inventory cycles’. *Review of Economic Statistics*, 23, 113–29.

Orphanides, A. and R. Solow (1990) ‘Money, inflation and growth’. In: B. Friedman and F. Hahn (eds): *Handbook of Monetary Economics*. Amsterdam: North Holland, 223–61.

Phillips, A.W. (1954) ‘Stabilisation Policy in a Closed Economy’. *The Economic Journal*, 64, 290–323.

Phillips, A.W. (1958) ‘The relation between unemployment and the rate of change of money wage rates in the United Kingdom, 1861–1957’. *Economica*, 25, 283–99.

Powell, A. and C. Murphy (1997) *Inside a Modern Macroeconometric Model. A Guide to the Murphy Model*. Heidelberg: Springer.

Roberts, J.M. (1997) ‘The wage curve and the Phillips curve’. Washington: Board of Governors of the Federal Reserve System.

Rose, H. (1967) ‘On the non-linear theory of the employment cycle’. *Review of Economic Studies*, 34, 153–73.

Rose, H. (1990) *Macroeconomic Dynamics. A Marshallian Synthesis.* Cambridge, Mass.: Basil Blackwell.

Rowthorn, B. (1980) *Capitalism, Conflict and Inflation.* London: Lawrence and Wishart.

Sargent, T. (1987) *Macroeconomic Theory.* New York: Academic Press.

Solow, R. and J. Stiglitz (1968) ‘Output, employment and wages in the short-run’. *Quarterly Journal of Economics*, 82, 537–60.

Solow, R. (1990) ‘Goodwin’s growth cycle: Reminiscence and rumination’. In: K.Velupillai (ed.): *Nonlinear and Multisectoral Macrodynamics*. London: Macmillan, 31–41.

Stock, J.H. and M.W. Watson (1997) ‘Business cycle fluctuations in US macroeconometric time series’. Harvard / Princeton University: Mimeo.

Tobin, J. (1975) ‘Keynesian models of recession and depression’. *American Economic Review*, 65, 195–202.

Turnovsky, S. (1977) *Macroeconomic Analysis and Stabilization Policies*. Cambridge: Cambridge University Press.

Turnovsky, S. (1995) *Methods of Macroeconomic Dynamics* Cambridge, MA: MIT Press.

# 7 Schumpeterian dynamics

## A disequilibrium theory of long run profits

*Katsuhito Iwai*

### 7.1 Introduction

The subtitle of this chapter may sound a contradiction in terms. In the traditional economic theory, by which I include both classical and neoclassical economics, the long-run state of an economy is an equilibrium state and the long-run profits (if they ever exist) are equilibrium phenomena. Figure 7.1 illustrates this by drawing two supply curves that can be found in any textbook of economics. In the upper panel is an upward-sloping supply curve which aggregates diverse cost conditions of the existing firms in an industry. Its intersection with a downward-sloping demand curve determines an equilibrium price, which in turn determines the amount of profits (represented by the shaded triangle) accruing to the industry as a whole. As long as the supply curve is upward-sloping, an industry is able to generate positive profits.

In traditional theory, however, this is merely a description of the 'short-run' state of an industry. Whenever there are positive profits, existing firms are encouraged to expand their productive capacities and potential firms are induced to enter the industry, both making the supply curve flatter and flatter. This process will continue until the industry supply curve becomes totally horizontal, thereby wiping out any opportunity for positive profits. The lower panel of Figure 7.1 describes this 'long-run' state of the industry.

This implies that if there are any profits in the long run, it must be the 'normal' profits which have already been incorporated into cost calculations. In fact, it is how to explain the fundamental determinants of these normal profits which divides the traditional economic theory into classical and neoclassical approaches. Classical economics (as well as Marxian economics) has highlighted an inverse relationship between the normal profit rate and the real wage rate, and reduced the problem of determining the former to that of determining the latter and ultimately to that of distributional conflicts between classes. Neoclassical economics has identified the normal profit rate with the interest rate plus a risk premium and reduced the problem of its determination to that of characterizing equilibrium conditions for intertemporal resource allocation under uncertainty. But no matter how opposed their views might appear over the ultimate determinants of normal profits, they

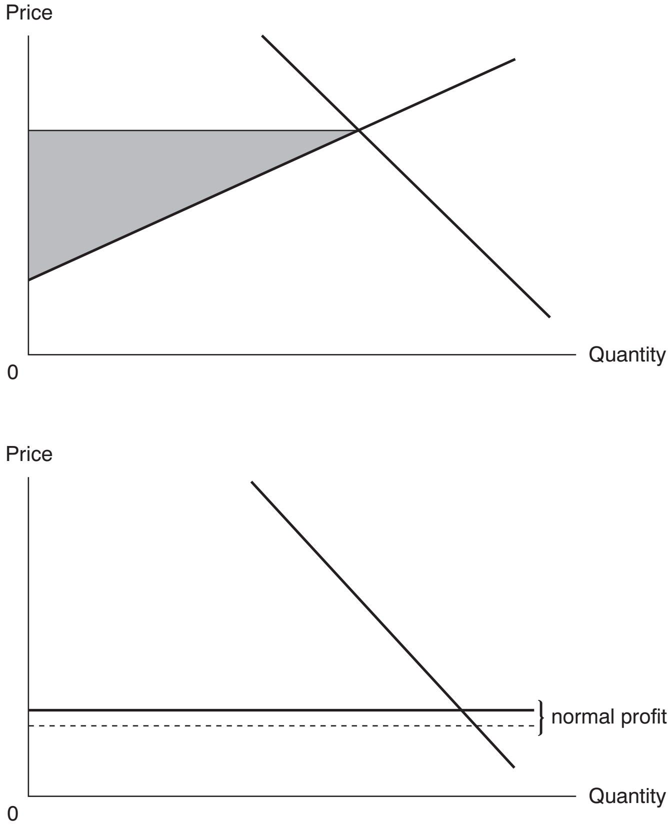


*Figure 7.1* Industry supply curve in the short run and in the long run

share the same ‘equilibrium’ perspective on long-run profits – any profits in excess of the normal rate are ‘disequilibrium’ phenomena which are bound to disappear in the long run.

It is Joseph Schumpeter who gave us a powerful alternative to this deep-rooted ‘equilibrium’ tradition in the theory of long-run profits. According to Schumpeter, it is through an ‘innovation’ or ‘doing things differently’ that positive profits emerge in the capitalist economy. ‘The introduction of new commodities . . . , the technological change in the production of commodities already in use, the opening-up of new markets or of new sources of supply, Taylorization of work, improved handling of material, the setting-up of new business organizations.’ (Schumpeter (1939), p. 84) etc. allow the innovators to charge prices much higher than costs of production. Profits are thus the premium put upon innovation. Of course, the innovator’s cost advantage does not last long. Once an innovation is successfully introduced into the economy, ‘it becomes much easier for other people to do the same thing’.[1] A subsequent wave of imitations soon renders the original innovation obsolete and gradually wears out the innovator's profit rate. In the long run, there is therefore an inevitable tendency towards classical or neoclassical equilibrium which does not allow any positive profits in excess of the normal rate. And yet Schumpeter argued that positive profits will never disappear from the economy because capitalism is 'not only never but never can be stationary'. It is an 'evolutionary process' that 'incessantly revolutionalizes the economic structure *from within*, incessantly destroying an old one, incessantly creating a new one'.[2] Indeed, it is to destroy the tendency towards classical or neoclassical equilibrium and to create a new industrial disequilibrium that is the function the capitalist economy has assigned to those who carry out innovations. 'Surplus values [i.e. profits in excess of normal rate] may be impossible in perfect equilibrium, but can be ever present because that equilibrium is never allowed to establish itself. They may always tend to vanish and yet be always there because they are incessantly recreated'.[3]

It is the first objective of this chapter to formalize this grand vision of Joseph Schumpeter from the perspective of evolutionary economics.[4] It makes use of a simple evolutionary model of Iwai (1984a, b) and demonstrates the Schumpeterian thesis that profits in excess of normal rate will never disappear from the economy no matter how long it is run. Indeed, it will be shown that what the economy will approach over a long passage of time is not a classical or neoclassical equilibrium of uniform technology but (at best) a statistical equilibrium of technological disequilibria which reproduces a relative dispersion of efficiencies among firms in a statistically balanced form. Although positive profits are impossible in perfect equilibrium, they can be ever present because that equilibrium is never allowed to establish itself.

This chapter is organized as follows. After having set up the static structure of an industry in Section 7.2, the following three sections will develop an evolutionary model of industrial dynamics and examine how the firms' capacity growth, technological imitations and technological innovations, respectively, move the industry's state of technology over time. It will be argued that while both the differential growth rates among different efficiency firms and the diffusion of better technologies through imitations push the state of technology towards uniformity, the punctuated appearance of technological innovations disrupts this equilibriating tendency. Section 7.6 will then turn to the long-run description of the industry's state of technology. It will indeed be shown that over a long passage of time these conflicting microscopic forces will balance each other in a statistical sense and give rise to a long-run distribution of relative efficiencies across firms. This long-run distribution will in turn allow us to deduce an *upward-sloping* long-run supply curve in Section 7.7. The industry is thus capable of generating positive profits even in the long run! Hence, the subtitle of this chapter – 'a disequilibrium theory of long-run profits'. Section 7.8 will then examine the factors determining the long-run profit rate of the industry.

The present chapter will adopt the 'satisficing' principle for the description of firms' behaviors – firms do not optimize a well-defined objective function but simply follow organizational routines in deciding their growth, imitation and innovation policies.[5] Indeed, the purpose of the penultimate Section 7.9 is to show that our evolutionary model is able to 'calibrate' all the macroscopic characteristics of the neoclassical growth model without having recourse to the neoclassical assumption of fully optimizing economic agents. If we look only at the aggregative performance of our evolutionary economy, it is as if aggregate labor and aggregate capital together produce aggregate output in accordance with a well-defined aggregate production function with Harrod-neutral technological progress. Yet, this macroscopic picture is a mere statistical illusion. If we zoomed into the microscopic level of the economy, what we would find is the complex and dynamic interactions among many a firm's capital growth, technological imitations and technological innovations. It is simply impossible to group these microscopic forces into a movement along an aggregate production function and a shift of that function itself. The neoclassical growth accounting may have no empirical content at all.

Section 7.10 concludes the chapter.

## 7.2 Construction of the industry supply curve

The starting point of our evolutionary model is an observation that knowledge is not a public good freely available among firms and that technologies with a wide range of efficiency coexist even in the same industry. And one of the end points of our evolutionary model is to demonstrate that technologies with a wide range of efficiency will indeed coexist even in the long-run.

Consider an industry which consists of many firms producing the same product.[6] Let us denote by $n$ the total number of technologies coexisting in the industry and assume that each of these technologies is of Leontief-type fixed-proportion technology with labor service as the sole variable input and capital stock as the sole fixed input. If we further assume that only labor productivity varies across technologies, we can express the $i_{\text{th}}$ technology as:

$$q = \min\left[\frac{l}{c_i}, \frac{k}{b}\right], \tag{7.1}$$

where $q, l$ and $k$ represent final output, labor input and capital stock, and $c_i$ and $b$ are labor and capital coefficients. Let us choose money wage as the numeraire. Then, the labor coefficient $c_i$ determines the unit cost of each technology up to a productive capacity $k/b$. I will slightly abuse the term and call $c_i$ the 'unit cost' of technology $i$. It is then possible to rearrange the indices of technology and let $c_n$ stand for the lowest and $c_1$ the highest unit cost of the industry without an loss of generality, or:

$$c_n < c_{n-1} < \ldots < c_i < \ldots < c_1. \tag{7.2}$$

I now have to introduce several notations in order to construct the supply curve of the industry in question. Let $k_t(c_i)$ represent the sum of the capital stocks of all the firms whose unit cost is $c_i$ at time $t$, and let $K_t(c_i) \equiv k_t(c_n) + \ldots + k_t(c_i)$ represent the cumulative sum of all the capital stocks of the firms whose unit costs are $c_i$ or lower at time $t$. The industry's total capital stock at time $t$ can then be represented

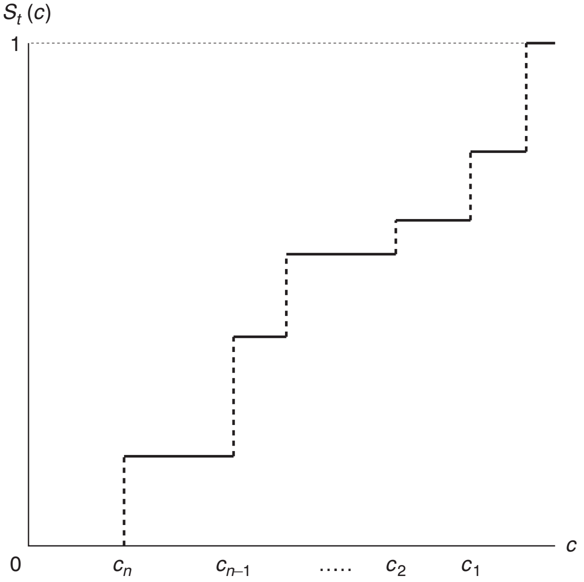


*Figure 7.2* Cumulative distribution of capacity shares

by $K_t(c_1)$, but will be denoted simply as $K_t$ in the following discussion. Next, let $s_t(c_i)$ and $S_t(c_i)$ represent the 'capital share' and the 'cumulative capital share' of a unit cost $c_i$ at time $t$. Of course, we have $s_t(c_i) \equiv k_t(c_i)/K_t$ and $S_t(c_i) \equiv K_t(c_i)/K_t$. As a convention, we set $S_t(c) = S_t(c_i)$ for $c_i \leq c < c_{i-1}$. Figure 7.2 exhibits a typical distribution of cumulative capital shares in the industry. It illustrates the 'state of technology' at a point in time by showing us how technologies with diverse unit costs are distributed among capital stocks of an industry.

The state of technology thus introduced, however, represents merely the production 'possibility' of an industry. How this possibility is actualized depends upon the price each firm is able to obtain in exchange for its product. Let us assume that the industry in question is a competitive industry in which a large number of firms are producing the same homogeneous product and charge the same price for it.[7] Let us denote by $P_t$ the product price (measured in terms of money wage) at time $t$. Then, under the assumptions of homogeneous product and fixed proportion technology, firms with unit costs strictly smaller than $P_t$ decide to produce up to their productive capacity $k/b$, and firms whose unit costs are strictly higher than $P_t$ decide to quit all production. Firms with the unit cost equal to $P_t$ are indifferent to their production level, as long as it does not exceed their productive capacity. (We ignore here the cost of shutting-down of a factory as well as the cost of setting-up of a new production line.)

It follows that when $c_{i-1} > P_t > c_i$ the total supply of the industry product becomes equal to $K_t(c_i)/b$ and that when $P_t = c_i$ it takes any value from $K_t(c_{i+1})/b$ to $K_t(c_i)/b$. Hence, if we denote by $Y_t(P)$ the industry's 'short-run supply curve'

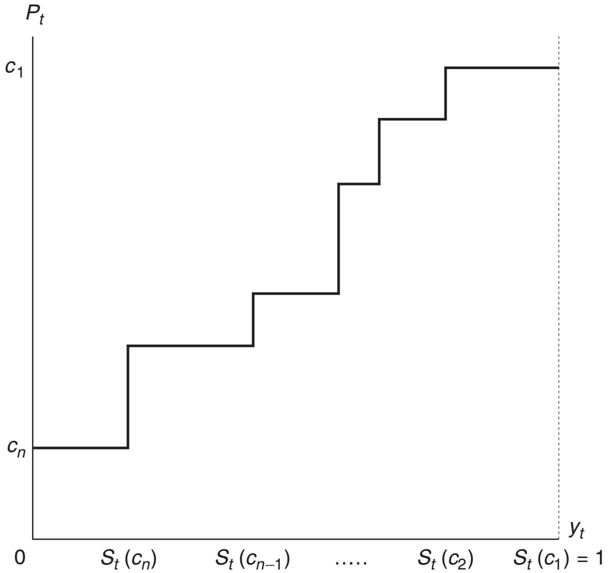


*Figure 7.3* Relative form of industry supply curve

(or short-run supply correspondence, to be precise ) at time $t$, it can be written as

$$\begin{aligned} Y_t(P) \quad &= K_t(c_i)/b & &\text{if } c_i < P < c_{i-1} \\ &\varepsilon[K_t(c_{i+1})/b, K_t(c_i)/b] & &\text{if } P = c_i. \end{aligned} \tag{7.3}$$

Dividing this by the total productive capacity $K_t/b$, we can also express it as:

$$\begin{aligned} y_t(P) \equiv Y_t(P)b/K_t \quad &= S_t(c_i) & &\text{if } c_i < P < c_{i-1} \\ &\varepsilon[S_t(c_{i+1}), S_t(c_i)] & &\text{if } P = c_i. \end{aligned} \tag{7.4}$$

(7.4) is nothing but the 'relative' form of industry supply curve at time $t$, which has neutralized the scale effect of changes in the total capital stock of the industry. Since the forces governing the motion of $S_t(c)$ are in general of different nature from those governing the motion of $K_t$, I will be concerned mostly with this relative form of industry supply curve in what follows.[8]

Figure7.3 depicts the relative form of industry supply curve, $y_t(P) \equiv Y_t(P)b/K_t$, in a Marshallian diagram with prices and costs (both in terms of money wage) measured along the vertical axis and quantities per unit of total productive capacity measured along the horizontal axis. Indeed, it merely turns Figure 7.2 around the 45 degree line. It is an *upward-sloping* curve as long as different unit costs coexist within the same industry.

## 7.3 Darwinian dynamics of the state of technology

Any freshman knows that the industry supply curve is a horizontal sum of all the individual supply curves existing in the industry. But the problem we now have to tackle is to ascertain how the dynamic competition among firms will mold the evolutionary pattern of the supply curve and govern the fate of the industry. This is not the problem for freshmen. Since there is a one-to-one correspondence between the relative form of industry supply curve and the cumulative distribution of capital shares, the analysis of the dynamic evolution of the former can be reduced to that of the latter.

Now, the state of technology in our Schumpeterian industry is moved by complex interactions among the dynamic forces working at the microscopic level of individual firms – successes and failures of technological innovations and imitations and the resulting differential growth rates among competing firms. Let us examine the effect of differential growth rates first.

'Without development there is no profit, without profit no development', so said our Schumpeter.[9] The following hypothesis relates the growth rate of capital stock to the rate of profit:

*Hypothesis (CG)*: The capital growth rate of a firm with unit cost $c_i$ is linearly increasing in its current rate of profit $r_t(c_i)$, or it is equal to:

$$\gamma r_t(c_i) - \gamma_0; \tag{7.5}$$

where $\gamma > 0$ and $\gamma_0 > 0$ are given constants. ■

This hypothesis needs little explanation. It merely says that a higher profit rate on the existing capital stocks stimulates capital accumulation, either by influencing the expected profitability of new investment projects or by directly providing an internal fund for the projects. The parameter $\gamma$ (or, more precisely, $\gamma/b$) represents the sensitivity of the firm's growth rate to the current profit rate, and the parameter $\gamma_0$ represents the rate of capital depreciation of the break-even firm. As I have already indicated in Section 7.1, the present chapter follows the strict evolutionary perspective in supposing that firms do not optimize but only 'satisfice' in the sense that they simply follow organizational routines in deciding their growth, imitation and innovation policies. Indeed, one of the purposes of this chapter is to see how far we can go in our description of the economy's dynamic performance without relying on the assumptions of individual optimality. I will therefore assume that the values of $\gamma$ and $\gamma_0$ are both exogenously given.[10]

We have already assumed that every firm in the industry produces the same homogeneous product and faces the same price $P_t$. If we further assume that the price of capital equipment is proportional to $P_t$, we can calculate the profit rate $r(c_i) \equiv (P_t y_t - c_i l_t)/P_t k_t$ as $b(P_t - c_i)/P_t$, which we will approximate as $b(\log P_t - \log c_i)$ for analytical convenience. Then, by simply differentiating the cumulative capacity share $S_t(c_i)$ with respect to time, *Hypothesis (CG)* allows us to deduce the following set of differential equations for the dynamics of cumulative capital shares[11]

$$\dot{S}_t(c_i) = \gamma\delta_t(c_i)S_t(c_i)(1 - S_t(c_i))(i = n, n-1, \ldots, 1). \tag{7.6}$$

In the above equations, $\delta_t(c_i)$ represents the difference between the logarithmic average of a set of unit costs higher than $c_i$ and the logarithmic average of a set of unit costs not higher than $c_i$, or:

$$\delta_t(c_i) \equiv \sum_{j=1}^{i-1} \frac{(\log c_j)s_t(c_j)}{1 - S_t(c_i)} - \sum_{j=i}^{n} \frac{(\log c_j)s_t(c_j)}{S_t(c_i)} > 0. \tag{7.7}$$

Its value in general depends on $t$ and the whole distribution of $c_i$. I will, however, proceed with the following analysis as if it were an exogenously given constant $\delta$, uniform both across technologies and over time. This will simplify the exposition of our evolutionary model immensely without losing any of its qualitative nature.[12] Then, we can rewrite (7.6) as:

$$\dot{S}_t(c_i) \cong \gamma\delta S_t(c_i)(1 - S_t(c_i)) \qquad (i = n, n-1, \ldots, 1). \tag{7.8}$$

Each of the above equations is a well-known 'logistic differential equation' with a logistic parameter $\mu$, and can be solved explicitly to yield:

$$S_t(c_i) = \frac{1}{1 + (1/S_T(c_i) - 1)e^{-\gamma\delta(t-T)}} \qquad (i = n, n-1, \ldots, 1), \tag{7.9}$$

where $e$ stands for the exponential and $T(\leq t)$ a given initial time.[13]
Differential growth rates among firms with different cost conditions never leave the industry's state of technology static. As the firms with relative cost advantage grow faster than the firms with relative cost disadvantage, the distribution of capital shares gradually shifts in favor of the lower unit costs, thereby reducing the average unit cost of the industry as a whole. This process then eliminates the relative cost advantage of the existing technologies one by one until the capital share of the least unit cost completely overwhelm those of the higher ones. Only the fittest will survive in the long run through their higher growth rates, and this of course is an economic analogue of the 'Darwinian' natural selection mechanism. The set of logistic equations (7.9) describes this 'economic selection' mechanism in the simplest possible mathematical form, and its evolutionary dynamics is illustrated by Figure 7.4. In particular, the equation for $i = n$ shows that the cumulative capital share of the lowest unit cost $S_t(c_n)$ moves along an $S$-shaped growth path. It grows almost exponentially when it occupies a negligible portion of the industry, gradually loses its growth momentum as its expansion narrows its own relative cost advantage, but never stops growing until it swallows the whole industry.

## 7.4 Lamarkian dynamics of the state of technology

Next, let us introduce the process of technological imitation and see how it molds the dynamics of the state of technology. In this chapter I will suppose that technology is not embodied in capital stocks and hypothesize the process of imitations as follows:[14]

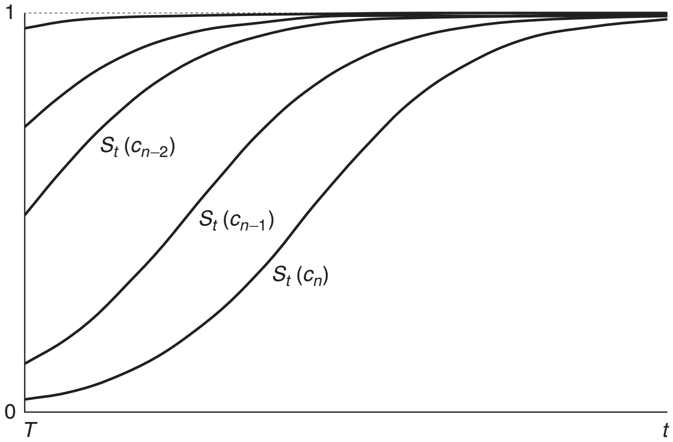


*Figure 7.4* Evolution of the state of technology under the pressure of either economic selection or technological diffusion

*Hypothesis* $(IM')$: The probability that a firm with unit cost $c_i$ succeeds in imitating a technology with unit cost $c_j$ is equal to:

$$\mu s_t(c_j)\mathrm{d}t \ \text{ if } c_j < c_i \text{ and } 0 \text{ if } c_j \geq c_i, \tag{7.10}$$

for a small time interval d$t$; where $\mu(> 0)$ is assumed to be a constant uniform across firms. ■

One of the characteristic features of technology is its non-excludability. It may be legally possible to assign property rights to the owners of technology. But, as Arrow has remarked in his classic paper (1962), 'no amount of legal protection can make a thoroughly appropriable commodity of something so intangible as information', because 'the very use of the information in any productive way is bound to reveal it, at least in part'.[15] The above hypothesis mathematically captures such spill-over effects of technology in the simplest possible manner. It says that it is much easier for a firm to imitate a technology with high visibility (i.e., a large capital share), than to imitate a technology with low visibility (i.e. a small capital share). Needless to say, the firm never imitates the technology whose unit cost is not smaller than its current one. The imitation coefficient $\mu$ in the above hypothesis represents the effectiveness of each firm's imitative activity. There is a huge body of literature, both theoretical and empirical, which identifies factors influencing the effectiveness of firms' imitation activities.[16] The main concern of the present chapter is, however, to work out the dynamic mechanism through which a given imitation policy of firms structures the evolutionary pattern of the industry's state of technology. In what follows I will simply assume that $\mu$ is an exogenously given parameter, uniform across firms and constant over time.[17]

In order to place the effect of technological diffusion in full relief, let us ignore the effect of economic selection for the time being. Then, the hypothesis *(IM')* allows us to deduce the following set of logistic differential equations as a description of the evolution of the state of technology under the sole pressure of technological imitations:[18]

$$\dot{S}_t(c_i) = \mu S_t(c_i)(1 - S_t(c_i)) \quad (i = n, n-1, \ldots, 1). \tag{7.11}$$

We have again encountered logistic differential equations, which can then be solved to yield the second set of logistic equations in this chapter!

$$S_t(c_i) = \frac{1}{1 + (1/S_T(c_i) - 1)e^{-\mu(t-T)}} \quad (i = n, n-1, \ldots, 1), \tag{7.12}$$

where $T(\leq t)$ is a given initial time.

Since the second set of logistic equations (7.12) is mathematically equivalent to the first set of logistic equations (7.9), Figure 7.4 in the preceding section can again serve to illustrate the dynamic evolution of the cumulative capacity shares under the sole pressure of technological diffusion. And yet, the logic behind these second logistic equations is entirely different from that of the first. 'If one or a few have advanced with success many of the difficulties disappear', so wrote Schumpeter, 'others can then follow these pioneers, as they will clearly do under the stimulus of the success now attainable. Their success again makes it easier, through the increasingly complete removal of the obstacles . . . , for more people follow suit, until finally the innovation becomes familiar and the acceptance of it a matter of free choice.'(Schumpeter (1961), p. 228) The logistic equations (7.12) describe this swarm-like appearance of technological imitations in the simplest possible form. In particular, the equation for $i = n$ shows that the cumulative capital share of the lowest unit cost moves along a $S$-shaped growth path, initially growing at an exponential rate but gradually decelerating its growth rate to approach unity asymptotically. In the long run, therefore, the lowest cost technology will dominate the whole industry, simply because it will eventually be diffused to all the firms in it. This technological diffusion process is nothing but an economic analogue of the 'Lamarkian' model of biological evolution – the achievement of one individual are passed directly to the other individuals.

Let us then bring back the Darwinian process of economic selection into our industry and add (7.6) to (7.11). The result is the third set of logistic differential equations in the present chapter:

$$\dot{S}_t(c_i) = (\gamma\delta + \mu)S_t(c_i)(1 - S_t(c_i)) \qquad (i = n, n-1, \ldots, 1), \tag{7.13}$$

which can again be solved explicitly as:

$$S_t(c_i) = \frac{1}{1 + (1/S_T(c_i) - 1)e^{-(\gamma\delta+\mu)(t-T)}} \qquad (i = n, n-1, \ldots, 1), \tag{7.14}$$

for $t \geq T$. (We refrain from drawing a diagram for the third set of logistic equations (7.14) which is qualitatively the same as Figure 7.4.)

We have thus shown how the mechanism of economic selection and the process of technological diffusion jointly contribute to the logistic growth process of cumulative capital shares – the former by amassing the industry's capacities in the hands of the lowest cost firms and the latter by diffusing the advantage of the lowest cost technology among imitating firms. While the former is Darwinian, the latter is Lamarkian. But, no matter how opposed the underlying logic might be, their effects upon the industry's state of technology are the same – the lowest cost technology will eventually dominate the whole capital stocks of the industry.

## 7.5 Punctuated dynamics of the state of technology

Does this mean that the industry's long-run state is no more than the paradigm of classical and neoclassical economics in which every market participant is supposed to have a complete access to the most efficient technology of the economy?

The answer is, however, 'No'. And the key to this negative answer lies, of course, in the phenomenon of innovation – the carrying out of what Schumpeter called a 'new combination'. Indeed, the functional role of innovative firms is precisely to destroy this tendency towards static equilibrium and to create a new industrial disequilibrium.

Suppose that at some point in time one of the firms succeeds in introducing a new technology with unit cost $c_{n+1}$ smaller than $c_n$. Let us denote this time by $T(c_{n+1})$ and call it the 'innovation time' for $c_{n+1}$. Then, a new cumulative capital share $S_t(c_{n+1})$ emerges out of nothing at $T(c_{n+1})$. Because of the disembodied nature of technology, $S_{T(c_{n+1})}(c_{n+1})$ is identical with the capital share of the innovator of $c_{n+1}$. Moreover, if the innovator's unit cost was, say, $c_i$ before innovation, all the cumulative capital shares from $S_t(c_{i+1})$ to $S_t(c_n)$ also experience a jump of the same magnitude at time $T(c_{n+1})$. In no time the innovator starts to expand its capital stocks rapidly, which then induces all the other firms to seek opportunities to imitate its technology. Through such selection mechanism and diffusion process, the newly created cumulative capital share begins to follow a $S$-shaped growth curve described by (7.14).

Innovation is not a single-shot phenomenon, however. No sooner than an innovation occurs, a new round of competition for a better technology begins. And no sooner than a new winner of this game is named, another round of technological competition is set out. The process repeats itself forever, and technologies with ever lower unit costs, $c_{n+2} > c_{n+3} > \ldots > c_N > \ldots$ will be introduced into an industry one by one at their respective innovation times $T(c_{n+2}), T(c_{n+3}), \ldots, T(c_N), \ldots$.

Figure7.5 shows how the industry's state of technology evolves over time, now as an outcome of the interplay among three dynamical forces working in the industry – economic selection mechanism, technological diffusion through imitations and creative destruction of innovations. In fact, while the former two work as equilibriating forces which tend the state of technology towards uniformity, the third works as a disequilibriating force which destroys this leveling tendency.

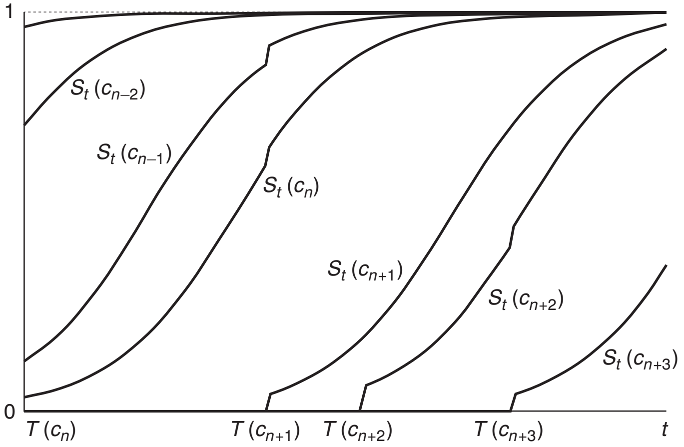


*Figure 7.5* Evolution of the state of technology under the joint pressure of economic selection, technological diffusion and recurrent innovations

A new question then arises: is it possible to derive any law-like properties out of this seemingly erratic movement of the industry state of technology?

In order to give an answer to this question, it is necessary to introduce two more hypotheses – one pertaining to invention and the other to innovation. The conceptual distinction between invention and innovation was very much emphasized by Schumpeter. Invention is a discovery of new technological possibility which is potentially applicable to the production processes of the economy. But, 'as long as they are not carried into practice', so says Schumpeter, 'inventions are economically irrelevant', and 'to carry any improvement into effect is a task entirely different from the inventing of it'.[19]

Denote then by $C(t)$ the unit cost of potentially the best possible technology at time $t$ and call it 'the potential unit cost'. The following is our hypothesis about the process of inventions:[20]

*Hypothesis (PC)*: The potential unit cost is declining at a positive constant rate $\lambda$ over time.

$$C(t) = e^{-\lambda t}, \tag{7.15}$$

where the scale of $C(0)$ is chosen to be unity. ■

The declining rate of potential unit cost $\lambda$ reflects the speed at which the stock of technological knowledge is being accumulated by academic institutions, private firms, government agencies and amateur inventors throughout the entire economy. In the present chapter which follows an evolutionary perspective, however, it is assumed to be given exogenously to the industry.

We are then able to characterize the notion of 'innovation' formally as an event in which the potential unit cost is put into actual use by one of the firms in the industry. This is tantamount to saying that when an innovation takes place at time $t$, it brings in a technology of unit cost $C(t)$ for the first time into an industry. This also implies that if a technology with unit cost $c$ is presently in use, it must have been introduced at time $t = T(c)$ where $T(c)$ is the inverse function of $C(t)$ defined by:

$$T(C(t)) \equiv t \text{ or } C(T(c)) \equiv c. \tag{7.16}$$

The function $T(c)$ thus defined is nothing but the 'innovation time' for unit cost $c$ we have already defined at the beginning of this section. Under the specification of the dynamics of potential unit cost in (7.15), we have $T(c) = -(\log c)/\lambda$.

Next, let us introduce the hypothesis about the process of innovations:

*Hypothesis* $(IN - a)$: The probability that a firm succeeds in an innovation is equal to:

$$v\mathrm{d}t, \tag{7.17}$$

during any small time interval d$t$, where $\nu$ is a small positive constant. ■

The parameter $\nu$ represents the effectiveness of each firm's innovative activity in the industry. Its value should in general reflect a particular innovation policy the firm has come to adopt in its long-run pursuit for technological superiority, and there is a huge body of literature identifying the factors which influence the firm's innovation policy.[21] Our main concern in the present article, however, is rather to examine how a given innovation policy will mold the evolutionary pattern of the industry's state of technology in the long run. In what follows I will simply assume that $\nu$ is an exogenously given parameter, uniform across firms and constant over time.[22] I will, however, suppose that the value of $\nu$ is much smaller than either that of $\mu$ or of $\gamma\delta$, for the innovations are by their nature much more difficult activities than imitations or growth.

Implicit in the above hypothesis is the supposition that an innovation can be introduced at any time and by any firm, irrespective of at what time and by which firm the last innovation was introduced.[23] Indeed, if we let $M$ denote the total number of firms in the industry, the probability that there is an innovation during a small time interval d$t$ is equal to $(\nu \mathrm{d}t)M = \nu M \mathrm{d}t$. Hence, the process of technological innovations in the industry as a whole constitute a Poisson process, which is sometimes called the law of rare events. As time goes by, however, innovations take place over and over again, and out of such repetitive occurrence of rare events a certain statistical regularity is expected to emerge.

## 7.6 The state of technology in the long run

Indeed, not only the process of innovations but also the entire evolutionary process of the state of technology is expected to exhibit a statistical regularity over a long passage of time. To see this, let $\hat{S}_t(c)$ denote the expected value of the cumulative capital share of $c$ at time $t$. For the purpose of describing the long-run pattern of the industry's state of technology, all we need to do is to follow the path of $\hat{S}_t(c)$. Indeed, it is not hard to deduce from *Hypothesis (IN-a)* the following set of differential equations for $\hat{S}_t(c)$:[24]:

$$\hat{S}_t(c) = (\gamma\delta + \mu)\hat{S}_t(c)(1 - \hat{S}_t(c)) + \nu(1 - \hat{S}_t(c)), \tag{7.18}$$

for $t \geq T(c)$. It turns out that this is the fourth set of logistic differential equations of this chapter, for each of which can be rewritten as $\dot{x} = (\gamma\delta + \mu + \nu)x(1 - x)$ with $x \equiv (\hat{S}_t(c) + \nu/(\gamma\delta + \mu))/(1 + \nu/(\gamma\delta + \mu))$. It can thus be solved to yield:

$$\hat{S}_t(c) = \frac{1 + \frac{\nu}{\gamma\delta+\mu}}{1 + \left(\frac{\gamma\delta+\mu}{\nu}\right) e^{-(\gamma\delta+\mu+\nu)(t-T(c))}} - \frac{\nu}{\gamma\delta + \mu}, \tag{7.19}$$

for $t \geq T(c)$.[25]

Of course, we cannot hope to detect any regularity just by looking at the motion of expected cumulative shares $\hat{S}_t(c)$ given above, for they are constantly pushed to the lower cost direction by recurrent innovations. If, however, we neutralize such declining tendency by measuring all unit costs $c$ relative to the potential unit cost $C(t)$ and observe the relative pattern of the cumulative capital shares, a certain regularity is going to emerge out of the seemingly unpredictable vicissitude of the industry's state of technology. Let us thus denote by $z$ the proportional gap between a given unit cost $c$ and the current potential unit cost $C(t)$, or

$$z \equiv \log c - \log C(t). \tag{7.20}$$

We call this variable the 'cost gap' of a given technology at time $t$. Since the inverse relationship between innovation time $T(c) = -(\log c)/\lambda$ and the potential unit cost $C(t) = \mathrm{e}^{-\lambda t}$ implies $z = \lambda(t - T(c))$, it is possible to rewrite (7.19) in terms of $z$ as follows:

$$\hat{S}_t(c) = \tilde{S}(z) \equiv \frac{1 + \alpha}{1 + (1/\alpha)e^{-\frac{1+\alpha}{\alpha\beta}z}} - \alpha, \tag{7.21}$$

where $\alpha$ and $\beta$ are composite parameters respectively defined by:

$$\alpha \equiv \frac{\nu}{\gamma\delta + \mu} \quad \text{and} \quad \beta \equiv \frac{\lambda}{\nu}. \tag{7.22}$$

This is the fifth time we have encountered a logistic curve. This time, it represents the 'long-run cumulative distribution' of cost gap $z$, towards which the relative form of the industry's state of technology has a tendency to approach in the long run. This distribution is a function only of the cost gap $z$ and is totally independent of calendar time $t$. Figure 7.6 illustrates this distribution.

As is seen from (7.21), the shape of $\tilde{S}(z)$ is determined completely by two composite parameters $\alpha$ and $\beta$, whose values are in turn determined through

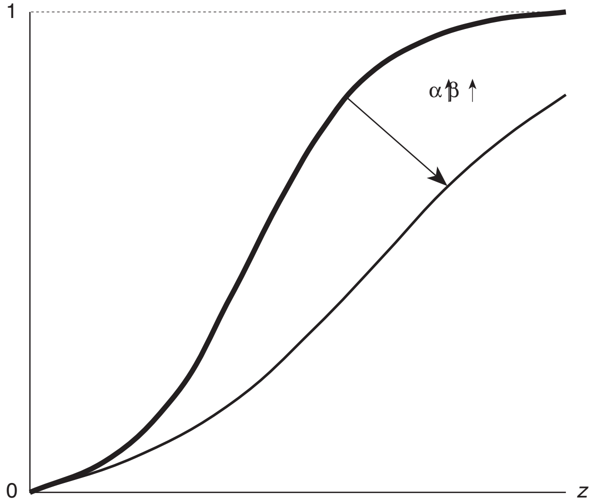


*Figure 7.6* Long-run cumulative distribution of capacity shares

(7.22) by such basic parameters as $\gamma\delta, \mu, \nu$ and $\lambda$. The first composite parameter $\alpha \equiv \nu/(\gamma\delta + \mu)$ represents the relative strength between the disequilibriating force of creative-cum-destructive innovations and the joint equilibriating force of economic selection mechanism and diffusion process through imitations. The second composite parameter $\beta \equiv \lambda/\nu$, on the other hand, represents the relative strength between the force of inventions and that of innovations. Since the expected rate of innovation per unit of time is $1/\nu$ and the reduction rate of the potential unit cost per unit of time is $\lambda, \beta$ can also be interpreted as the expected cost reduction rate of each innovation.[26] It is not difficult to show that:[27]

$$\frac{\partial \tilde{S}(z)}{\partial \alpha} < 0 \quad \text{and} \quad \frac{\partial \tilde{S}(z)}{\partial \beta} < 0. \tag{7.23}$$

As is illustrated in Figure 7.6, an increase in both $\alpha$ and $\beta$ thus shifts $\tilde{S}(z)$ clockwise, thus rendering the distribution of efficiencies across firms more disperse than before.

The long-run cumulative distribution $\tilde{S}(z)$ thus deduced is a statistical summary of the way in which a multitude of technologies with diverse cost conditions are dispersed among all the existing capital stocks of the industry. It shows that, while the on-going inventive activities are constantly reducing the potential unit cost, the unit costs of a majority of production methods actually in use lag far behind this potential one. The state of technology therefore has no tendency to approach a classical or neoclassical equilibrium of uniform technology even in the long run. What it approaches over a long period of time is merely a ‘statistical equilibrium of technological disequilibria’.

## 7.7 The industry supply curve in the long run

Now, the fact that the state of technology retains the features of disequilibrium even in the long run does have an important implication for the nature of the industry's long-run supply curve. For, as is seen by (7.4), the relative form of industry supply curve $y_t = Y_t(P_t)b/K_t$ traces the shape of $S_t(c)$, except for the portions of discontinuous jumps. Hence, if the expectation of $S_t(c)$ tends to exhibit a statistical regularity in the form of $\tilde{S}(z)$, the expectation of the relative form of the industry supply curve should also exhibit a statistical regularity in the same long-run form of $\tilde{S}(z)$. Let us denote by $p_t$ the relative gap between a given product price $P_t$ and the potential unit cost $C(t)$, or

$$p_t \equiv \log P_t - \log C(t), \tag{7.24}$$

and call it the 'price gap' at time $t$. Then, we can obtain the following proposition without paying any extra cost.

*Proposition (SC)*: Under *Hypotheses (CG), (IM'), (PC)* and *(IN-a)*, the expected value of the relative supply curve of the industry $y_t = Q_t(P_t)b/K_t$ will in the long run approach a functional form of

$$\tilde{S}(p_t) \equiv \frac{1+\alpha}{1+(1/\alpha)e^{-\frac{1+\alpha}{\alpha\beta}p_t}} - \alpha. \qquad \blacksquare \tag{7.25}$$

Figure 7.7 exhibits the relative form of the industry's long-run supply curve as a function of price gap $p$. As a matter of fact, it has been drawn simply by turning Figure 7.6 around the 45 degree line. It therefore moves counter-clockwise as either of the composite parameters $\alpha$ and $\beta$ increases. This implies that the long-run supply curve becomes more upward-sloping, as the disequilibriating force of creative-cum-destructive innovations becomes stronger than the joint equilibriating force of economic selection and technological diffusion or as the average rate of cost reduction of each innovation becomes larger.

What is most striking about this long-run supply curve, however, is not that it is the 'sixth' logistic curve we have encountered in this chapter but that it is an *upward-sloping* supply curve!

Let us recall the lower panel of Figure 7.1 of the introductory section. It reproduced a typical shape of the long-run supply curve which can be found in any textbook of economics. This horizontal curve was supposed to describe the long-run state of the industry in which the least cost technology is available to every firm in the industry and all the opportunities for positive profits are completely wiped out. However, the relative form of the long-run supply curve we have drawn in Figure 7.7 has nothing to do with such traditional picture. There will always be a multitude of diverse technologies with different cost conditions, and the industry supply curve will never lose an upward-sloping tendency, just as in the case of the 'short-run' supply curve of the upper panel of Figure 7.1. There are, therefore, always some firms which are capable of earning positive profits, no matter how competitive the industry is and no matter how long it is run.

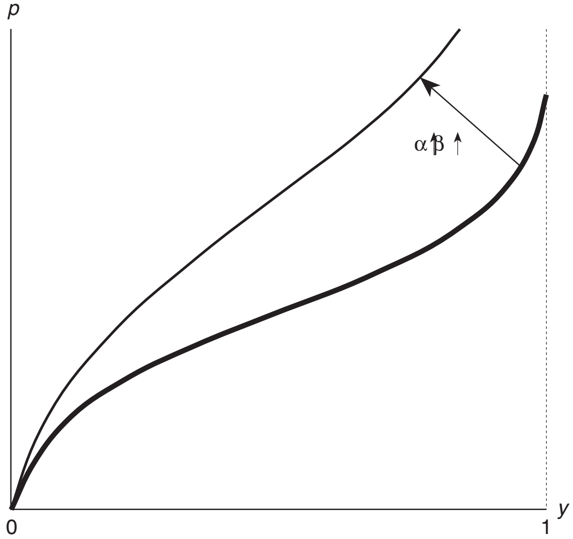


*Figure 7.7* Long-run industry supply curve

We can thus conclude that positive profits are not only the short-run phenomenon but also the long-run phenomenon of our Schumpeterian industry. It is true that the positivity of profits is a symptom of disequilibrium. But, if the industry will approach only a statistical equilibrium of technological disequilibria, it will never stop generating positive profits from within even in the never-never-land of long-run.

## 7.8 The determination of the long-run profit rate

It is one thing to demonstrate the existence of positive profits in the long run. It is, however, another to analyze the factors which determine the long-run profit rate.

Let us then look at Figure 7.8 which superimposes a demand curve on Figure 7.7. If we suppose that this demand curve is shifting to the right at the same rate as that of the industry's total capital stock and shifting to the bottom at the same rate as that of the potential unit cost, its relative form will become invariant over time. The intersection $e^*$ of this relative demand curve with the long-run relative supply curve then determines the long-run equilibrium price gap $p^*$ and the long-run equilibrium output-capacity ratio $y^* = \tilde{S}(p^*)$. Since we have approximated the profit rate $(Pq - cl)/Pk$ of each technology by $b(\log P - \log c)$, we can also express it as $b((\log P - C(t)) - (\log c - C(t))) = b(p - z)$. This is nothing but the vertical distance between a given price gap and a point on the upward-sloping supply curve. Summing these individual profit rates from $z = 0$ to $z = p^*$ with capital shares $\tilde{s}(z) \equiv \tilde{S}'(z)$ as relative weights, we can calculate the long-run profit

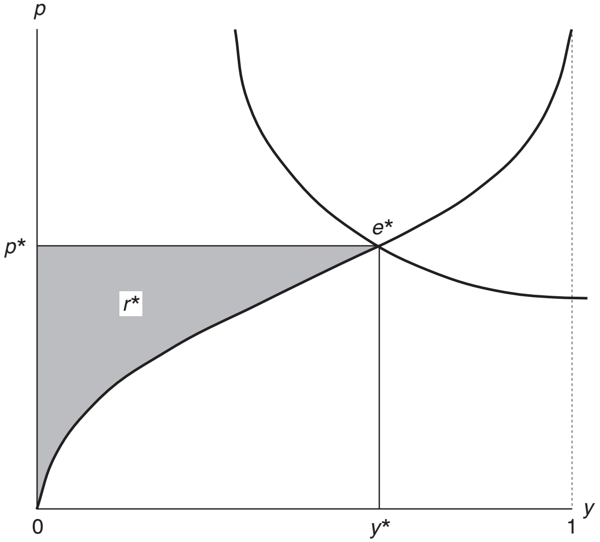


*Figure 7.8* Determination of long-run profit rate

rate $r^*$ of the industry as a whole. Graphically, it can be represented by the shaded area $0e^*y^*$ in Figure 7.8. Algebraically, it can be expressed as:[28]

$$\begin{aligned} r^* &= \int_0^{p^*} b(p^*-z)\tilde{s}(z)\mathrm{d}z = b\int_0^{p^*}\tilde{S}(z)\mathrm{d}z \\ &= \left(\frac{\alpha\beta b}{1+\alpha}\right)\left(-\log(1-y^*) - \alpha\log\left(1+\frac{y^*}{\alpha}\right)\right) > 0. \end{aligned} \tag{7.26}$$

We have thus succeeded in giving a complete characterization of the long-run profit rate of our Schumpeterian industry. It is *positive*, indeed.

It is now time to do some comparative dynamics. First, demand effects. It does not require any graphical explanation to see that an upward shift of the relative demand curve works to increase the long-run profit rate of the industry $r^*$. In fact, a differentiation of (7.26) with respect to $y^*$ leads to:

$$\frac{\partial r^*}{\partial y^*} = \frac{\alpha\beta b}{1+alpha}\left(\frac{1}{1-y^*} - \frac{\alpha}{\alpha+y^*}\right) > 0. \tag{7.27}$$

A further differentiation of (7.27) leads to:

$$\frac{\partial^2 r^*}{\partial y^{*2}} = \frac{\alpha\beta b}{1+\alpha}\left(\frac{1}{(1-y^*)^2} + \frac{\alpha^2}{(\alpha+y^*)^2}\right) > 0. \tag{7.28}$$

The industry's long-run profit rate $r^*$ is thus seen to be an increasing and convex function of the equilibrium output-capacity ratio $y^*$.

This convex relationship between long-run profit rate and output–capital ratio would have a particularly important implication for the dynamic stability, or more appropriately, dynamic instability of our Schumpeterian economy. For *Hypothesis (CG)* immediately implies that the growth rate of fixed investment also becomes on average an increasing and convex function of output–capacity ratio, which is very likely to violate the stability condition for investment-saving equilibrium of the economy as a whole. In the present chapter, however, we can only mention this possibility in passing and must resume our comparative dynamics.

Next, let us examine the effects of a shift in the long-run supply curve on the industry's long-run profit rate. This, however, turns out to be a much more involved exercise than that on the demand effects. I will therefore relegate the detailed discussions to Appendix and only summarize the results obtained therein.

When the relative demand curve is perfectly elastic with respect to price change, we have:

$$\left.\frac{\partial r^*}{\partial \alpha}\right|_{p^* = \text{const.}} < 0 \text{ and } \left.\frac{\partial r^*}{\partial \beta}\right|_{p^* = \text{const.}} < 0. \tag{7.29}$$

In this case, both an intensification of the force of innovations relative to the force of growth and imitations and an intensification of the force of invention relative to that of innovations reduce the profit rate of the industry in the long run. But the assumption of perfectly elastic industry demand curve is empirically of limited relevancy (except for the case of price regulation), and we had better proceed to another special case.

When the relative demand curve is absolutely inelastic with respect to price change, we have:

$$\left.\frac{\partial r^*}{\partial \alpha}\right|_{y^* = \text{const.}} > 0, \text{ and } \left.\frac{\partial r^*}{\partial \beta}\right|_{y^* = \text{const.}} > 0. \tag{7.30}$$

In this case, as the disequilibriating force of creative-cum-destructive innovations becomes stronger than the equilibriating force of economic selection and swarm-like imitations, or as the average rate of cost reduction of each innovation becomes greater, the industry is expected to generate a higher profit rate in the long-run. Innovation is not only the source of short-run profits but also the source of long-run profits in an industry with inelastic demand.

Finally, when the relative demand is neither perfectly elastic nor absolutely inelastic, we have:

$$\frac{\partial r^*}{\partial \alpha} > (<) 0 \text{ and } \frac{\partial r^*}{\partial \beta} > (<) 0, \tag{7.31}$$

when demand curve is inelastic (elastic).

In this general case, as the disequilibriating force of creative-cum-destructive innovations becomes stronger than the equilibriating forces of economic selection and swarm-like imitations, or as the average rate of cost reduction of each innovation becomes greater, the industry is expected to generate a higher profit rate in the long run, as long as the price-elasticity of the demand curve is not so large. However, this tendency will be reversed when the industry demand curve becomes sufficiently elastic with respect to price change.

## 7.9 *Pseudo*-aggregate production functions

Since the pioneering works of Solow (1956, 1957), it has become the standard method of neoclassical economics to use the concept of an ‘aggregate production function’ in accounting the sources of economic growth. It allows economists to decompose variations in GNP into those due to movements along the aggregate production function and those due to shifts of the aggregate production function itself. The former can be attributed to changes in measurable inputs, usually capital and labor, and the latter to changes in technology, an unobservable variable usually inferred from the data as a residual. Early empirical studies of the long-term macroeconomic growth in advanced capitalist economies found that only a very small portion of GNP growth can be accounted for by increases in capital and labor and that most of the growth being explained by technological progress – an increase in the residual factor. More recent efforts by Maddison (1987) and others, however, have succeeded in reducing the size of the residual factor substantially by incorporating variations in the qualities of capital and labor and other supplementary effects.

The ‘success’ of the neoclassical growth accounting exercises is quite impressive. The challenge to any theory claiming to challenge the neoclassical orthodoxy is therefore to match its power of tracking down the empirical patterns of the macroscopic growth processes of advanced capitalist economies. The most straightforward way to do this is, of course, to set up an empirical study of our own. But in order not to lengthen this already lengthy chapter, I choose a short-cut. The purpose of this section is to demonstrate that our evolutionary model is capable of ‘calibrating’ all the characteristics of a neoclassical aggregate production function both in the short run and in the long run.[29] If the neoclassical growth model is capable of accounting the actual macroeconomic growth paths of advanced capitalist economies, then our evolutionary model is equally capable of performing the same task. There is no way to differentiate these two models empirically at the macroscopic level. Moreover, our evolutionary model has a decided advantage over the neoclassical model in its ability to integrate microeconomic processes with macroeconomic phenomena. While the neoclassical growth theory simply ignores the complexity of the growth processes we daily observe at the microscopic level, its recognition is the very starting point of our evolutionary model.

Let me begin this ‘calibration’ exercise by computing the amount of labor employment for each level of product demand. When product demand is small so that price $P_t$ just covers the minimum unit cost $c_n$, only the first-best technology firms can engage in production and the level of product demand determines that of output $Y_t$. Because of the fixed proportion technology (7.1), the level of total employment $L_t$ associated with this output is $c_n Y_t$. When the demand reaches the total capacity of the best technology $k_t(c_n)/b = s_t(c_n)K_t/b$, a further increase in demand is absorbed solely by an increase in $P_t$, while output is kept at the capacity level. But when $P_t$ reaches $c_{n-1}$, the second-best technology firms start to produce and all the increase in demand is absorbed by a corresponding increase in output. The relation between output and employment can then be given by $L_t = c_n s_t(c_n)K_t/b + c_{n-1}(Y_t - s_t(c_n)K_t/b)$ until $Y_t$ reaches the total productive capacity of the first- and second-best technology firms $(s_t(c_n) + s_t(c_{n-1}))K_t/b$. In general, the relation between $Y_t$ and $L_t$ can be given by

$$\begin{aligned} L_t &= \sum_{j=n}^{j=i} c_j s_t(c_j)K_t/b + c_{i-1}\left(Y_t - \sum_{j=n}^{j=i} s_t(c_j)K_t/b\right) \\ &\equiv \int_0^{c_i} c\mathrm{d}S_t(c)K_t/b + c_{i-1}(Y_t - S_t(c_i)K_t/b) \end{aligned}$$

whenever $S(c_i)K_t/b \leq Y_t < S(c_{i-1})K_t/b$. If we divide this relation by $K_t/b$ and take its inverse, we can construct a functional relation between the industry-wide labor–capacity ratio $l_t \equiv L_t b/K_t$ and the industry-wide output–capacity ratio $y_t \equiv Y_t b/K_t$ as:

$$y_t = f_t(l_t), \tag{7.32}$$

where $l \equiv \int_0^{c_i} c\mathrm{d}S_t(c) + c_{i-1}(f_t(l) - S_t(c_i))$ whenever $S_t(c_i) \leq f_t(l) < S_t(c_{i-1})$. Figure 7.9 depicts this functional relation in a Cartesian diagram which measures labor–capacity ratio $l$ along horizontal axis and output-capacity ratio $y$ along a vertical axis. It is evident that this relation satisfies all the properties a neoclassical production function is supposed to satisfy.[30] $Y$ is linearly homogeneous in $L$ and $K$, because $y \equiv Yb/K$ is a function only of $l \equiv Lb/K$. Though not smooth, this relation also allows a substitution between $K_t$ and $L_t$ and satisfies the marginal productivity principle: $\partial\bar{y}_t/\partial l_t \leq 1/P_t \leq \partial^+ y_t/\partial l_t$. (Here, $1/P_t$ represents a real wage rate because of our choice of money wage rate as the numeraire, and $\partial\bar{y}/\partial l$ and $\partial^+ y/\partial l$ represent left- and right-partial differential, respectively.) Yet, the important point is that this is not a production function in the proper sense of the word. It is a mere theoretical construct that has little to do with the actual technological conditions of the individual firms working in the industry. It is in this sense that we call the relation (7.32) a ‘short-run *pseudo*-aggregate production function’, with an emphasis on the adjective: ‘*pseudo*’.

The shape of the short-run *pseudo*-production function $y = f_t(l)$ is determined by a distribution of capital shares $\{S_t(c_i)\}$ across technologies. Hence, as this distribution changes, the shape of this short-run function also changes. And in our Schumpeterian industry, the distribution of capital shares is incessantly changing over time as the result of dynamic interplay among capital growth, technological innovation and technological imitation. The most conspicuous feature of the short-run *pseudo*-production function is, therefore, its instability.

In the long run, however, we know we can detect a certain statistical regularity in the distribution of capital shares $\{S_t(c_i)\}$ out of its seemingly unpredictable

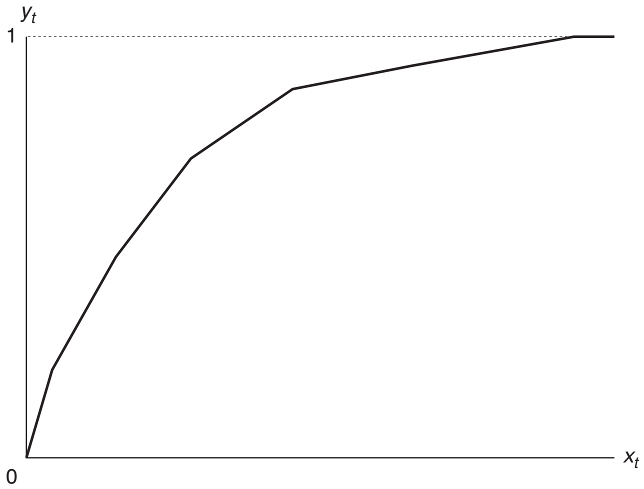


*Figure 7.9* Short-run *pseudo*-aggregate production function

movement. We can thus expect to detect a certain statistical regularity in the *pseudo*-production function as well out of its seemingly unpredictable movement. Let $\hat{l}$ and $\hat{y}$ denote the expectation of labor–capacity ratio $l \equiv Lb/K$ and of output–capacity ratio $y \equiv Yb/K$, respectively. Then, we indeed arrive at:[31]

*Proposition (PF)*: Under *Hypotheses (CG), (IM'), (PC)* and *(IN-a)*, the functional relationship between the expected labor–capacity ratio and the expected output–capacity ratio will in the long run take the form of :

$$\hat{y} = \tilde{f}\left(\hat{l}e^{\lambda t}\right), \tag{7.33}$$

where the function $\tilde{f}(\cdot)$ is defined implicitly by the following identity:

$$le^{\lambda t} \equiv \int_0^{\tilde{f}\left(le^{\lambda t}\right)} \left(\frac{\alpha + y}{\alpha(1-y)}\right)^{\frac{\beta\alpha}{1+\alpha}} \mathrm{d}y. \qquad \blacksquare \tag{7.34}$$

At the seventh time, we have finally graduated from the tyranny of logistic equations! What we have obtained in Figure 7.10 is a well-behaved function which satisfies all the characteristics a neoclassical production function should have. Indeed, it is not hard to show that $\tilde{f}(0) = 0, \tilde{f}'(\cdot) > 0, \tilde{f}''(\cdot) < 0$.[32] It is as if total labor force $L$ and total capital stock $K$ produce the total output $Y$ in accordance with an aggregate neoclassical production function $\tilde{f}(\cdot)$ with pure labor augmenting (or Harrod-neutral) technological progress $e^{\lambda t}$. It is, in other words, as if we had entered the Solovian world of neoclassical economic growth where the economy's macroscopic growth process could be decomposed into the capital–labor substitution along an aggregate neoclassical production function and the constant outward shift of the aggregate neoclassical production function itself. This is, however, a

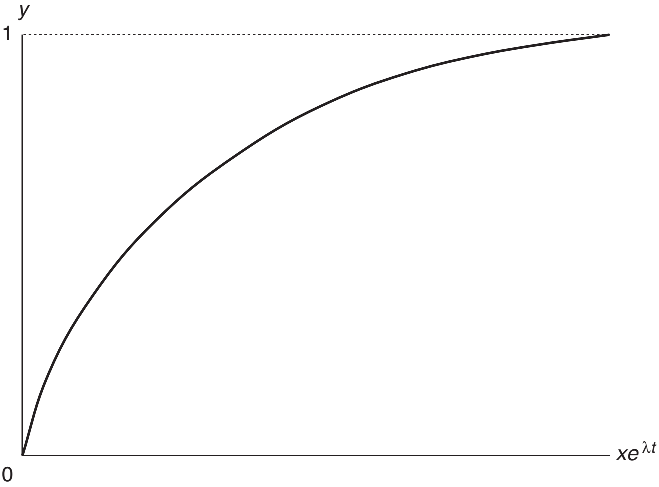


*Figure 7.10* Long-run *pseudo*-aggregate production function

mere statistical illusion! If we zoomed into the microscopic level of the economy, what we would find is the complex and dynamic interactions among many a firm's capital growth, technological imitation and technological innovation. In fact, as is seen from (7.34), the functional form of $\tilde{f}(\cdot)$ is a complex amalgam of such basic parameters of our Schumpeterian model as $\alpha \equiv \nu/(\gamma\delta + \mu)$ and $\beta \equiv \lambda/\nu$. It is just impossible to disentangle various microscopic forces represented by these parameters and decompose the overall growth process into a movement along a well-defined aggregate production function and an outward shift of the function itself.[33] Indeed, it is not hard to show that both an increase in $\alpha$ and in $\beta$ shift the function $\tilde{f}(\cdot)$ in the downward direction,[34] or

$$\frac{\partial \tilde{f}(\cdot)}{\partial \alpha} < 0 \text{ and } \frac{\partial \tilde{f}(\cdot)}{\partial \beta} < 0. \tag{7.35}$$

We are after all living in a Schumpeterian world where the incessant reproduction of technological disequilibria prevents the aggregate relation between capital and labor from collapsing into the fixed proportion technology of individual firms. It is, in other words, its non-neoclassical features that give rise to the macroscopic illusion that the industry is behaving like a neoclassical growth model. It is for this reason we will call the relation (7.33) the 'long-run *pseudo* aggregate production function'.

## 7.10 Concluding remarks

In the traditional economic theory, whether classical or neoclassical, the long-run state of the economy is an equilibrium state and the long-run profits are equilibrium phenomena. If there is a theory of long-run profits, it must be a theory about the determination of the normal rate of profit.

This chapter has challenged this long-held tradition in economics. It has introduced a simple evolutionary model which is capable of analyzing the development of the industry's state of technology as a dynamic interplay among many a firm's growth, imitation and innovation activities. And it has demonstrated that what the industry will approach over a long passage of time is not a classical or neoclassical equilibrium of uniform technology but a statistical equilibrium of technological disequilibria which maintains a relative dispersion of efficiencies in a statistically balanced form. Positive profits will never disappear from the economy no matter how long it is run. 'Disequilibrium' theory of 'long-run profits' is by no means a contradiction in terms.

Not only is a disequilibrium theory of long-run profits possible, but it is also 'operational'. Indeed, our evolutionary model has allowed us to calculate (only with pencils and paper) the economy's long-run profit rate as an explicit function of the model's basic parameters representing the microscopic forces of capital growth, technological imitations, recurrent innovations and steady inventions. 'Without development there is no profit, without profit no development', to quote Joseph Schumpeter once more.[35] The model we have presented in this chapter can thus serve as a foundation, or at least as a building block, of the theory of 'long-run development through short-run fluctuations' or 'growth through cycles'. To work out such a theory in more detail is of course an agenda for the future research.

## Appendix: Comparative dynamics of supply-side determinants of long-run profit rate

The purpose of this Appendix is to deduce (7.29), (7.30) and (7.31).

Consider first the case of perfectly elastic demand curve. Although the economic relevancy of this special case is of limited nature, it serves as a useful benchmark for the other cases. Figure 7.A1 juxtaposes a horizontal demand curve on the relative form of a long-run supply curve. We already know from Section 7.8 that an increase in either $\alpha$ or $\beta$ moves the inverted logistic shape of the supply curve counter-clockwise. As is seen from Figure 7.A1, such a supply curve shift transfers the equilibrium point from $e^*$ to $e^{**}$ along the horizontal demand curve and squeezes the long-run profit rate by the magnitude equal to $A \equiv 0e^*e^{**}$. We can easily confirm this graphical exposition by differentiating (7.26) with respect to $\alpha$ and $\beta$, keeping $p^*$ constant.

$$\begin{aligned}
\left.\frac{\partial r^*}{\partial \alpha}\right|_{p^*=\text{ const.}} &= b\int_0^{p^*} \frac{\partial \tilde{S}(z)}{\partial \alpha}\mathrm{d}z \equiv -A_\alpha \\
&= -\frac{\beta b}{\alpha(1+\alpha)}\int_{e^{-(1+y^*)/(\alpha\beta)}}^{1} \frac{u-1-\log u}{(1+u)^2}\mathrm{d}u < 0; \\
\left.\frac{\partial r^*}{\partial \beta}\right|_{p^*=\text{ const.}} &= b\int_0^{p^*} \frac{\partial \tilde{S}(z)}{\partial \beta}\mathrm{d}z \equiv -A_\beta \\
&= -\frac{(1+\beta)b}{\alpha\beta}\int_{e^{-(1+y^*)/(\alpha\beta)}}^{1} \frac{u-1-\log u}{(1+u)^2}\mathrm{d}u < 0.
\end{aligned} \tag{7.A1}$$

This is nothing but (7.29) of the main text. Note also that since $\partial\tilde{S}(z)/\partial\nu > 0$, $\frac{\partial r^*}{\partial\nu}\big|_{p^*=\text{ const.}} > 0$.

Next, consider the case of an absolutely inelastic demand curve. As is shown in Figure 7.A2 which juxtaposes a vertical demand curve on an inverted logistic shape of the long-run supply curve, an increase in either $\alpha$ or $\beta$ moves the latter counter-clockwise and transfers the equilibrium point from $e^*$ to $e^{**}$ along the vertical demand curve. This raises $p^*$ to $p^{**}$, while keeping $y^*$ the same as before. The long-run profit rate thus changes from $0e^*p^*$ to $0e^{**}p^{**}$. We have to examine whether this amounts to an increase or decrease of $r^*$. To see this, Figure 7.A2 decomposes this change of profit rate into two components – $A \equiv 0e^*e^{**}$ and $B \equiv p^*e^*e^{**}p^{**}$. The first component $A$ represents the 'loss' of profit rate due to a universal increase of cost gaps, which corresponds to the profit loss $A$ of the previous case. In the present case of absolutely inelastic demand curve, however, an increase in the long-run equilibrium price gap gives rise to a 'gain' of profit rate, as is represented by the second component $B$. Whether $r^*$ increases or decreases thus depends on whether $A$ is smaller or larger than $B$. This can be checked by differentiating (7.26) with respect to $\alpha$ and $\beta$, keeping $y^*$ constant. We have:

$$\begin{aligned}
\left.\frac{\partial r^*}{\partial\alpha}\right|_{y=\text{ const.}} &= b\int_0^{\tilde{S}-1(y^*)}\frac{\partial\tilde{S}(z)}{\partial\alpha}\mathrm{d}z + by^*\frac{\partial\tilde{S}^{-1}(y^*)}{\partial\alpha}\\
&\equiv -A_\alpha + B_\alpha.\\
&= \frac{\beta b}{(1+\alpha)^2}\left(\left(-\log(1-y^*)-\alpha\log\left(1+\frac{y^*}{\alpha}\right)\right)\right.\\
&\quad\left.-(1+\alpha)\alpha\left(\log\left(1+\frac{y^*}{\alpha}\right)-\frac{y^*}{\alpha+y^*}\right)\right); \qquad \text{(7.A2)}\\
\left.\frac{\partial r^*}{\partial\beta}\right|_{y^*=\text{ const.}} &= b\int_0^{\tilde{S}^{-1}(y^*)}\frac{\partial\tilde{S}(z)}{\partial\beta}\mathrm{d}z + by^*\frac{\partial\tilde{S}^{-1}(y^*)}{\partial\beta} \equiv -A_\beta + B_\beta\\
&= -\left(\frac{\alpha b}{1+\alpha}\right)\left(\log(1-y^*)+\alpha\log\left(1+\frac{y^*}{\alpha}\right)\right) > 0.
\end{aligned}$$

Although both $-\log(1-y^*)-\alpha\log(1+\frac{y^*}{\alpha})$ and $(1+\alpha)\alpha(\log(1+\frac{y^*}{\alpha})-\frac{y^*}{\alpha+y^*})$ are positive in the first expression, the former dominates the latter if we let $\alpha\to 0$. Since $\alpha\equiv\nu/(\gamma\delta+\mu)$ is assumed to be small, it does not seem unreasonable to suppose the first expression to be positive. The second expression is always positive. Hence, (7.30) of the main text. Note that we can also calculate $\frac{\partial r^*}{\partial\nu}\big|_{y^*=\text{const.}}$ as

$$\frac{-\alpha^2\beta b}{(1+\alpha)^2}\left(\left(-\log(1-y^*)-\frac{\alpha y^*}{\alpha+y^*}\right)-\left(\log\left(1+\frac{y^*}{\alpha}\right)-\frac{y^*}{\alpha+y^*}\right)\right)<0.$$

Finally, let us consider the general case where industry demand curve is neither perfectly elastic nor absolutely inelastic. As is seen from Figure 7.A3, an increase in either $\alpha$ or $\beta$ transfers the equilibrium point upward from $e^*$ to $e^{**}$ along this downward-sloping demand curve. This raises $p^*$ to $p^{**}$ but lowers $y^*$ to $y^{**}$, thereby changing $r^*$ from $0e^*y^*$ to $0e^{**}y^{**}$. We can then decompose this change again into $A\equiv 0e^*e^{**}$ and $B'\equiv p^*e^*e^{**}p^{**}$. $A$ represents the 'loss' of $r^*$ due to a universal

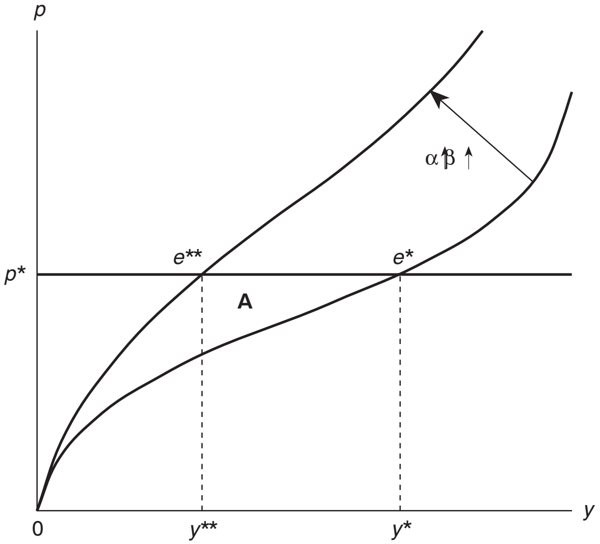


*Figure 7.A1* The case of a perfectly elastic demand curve

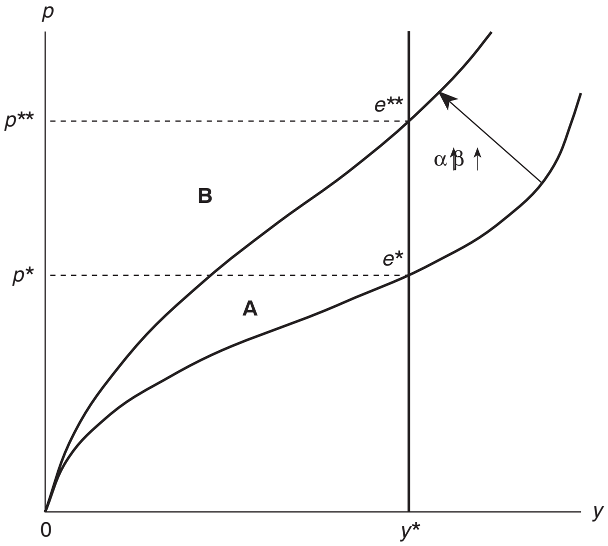


*Figure 7.A2* The case of an absolutely inelastic demand curve

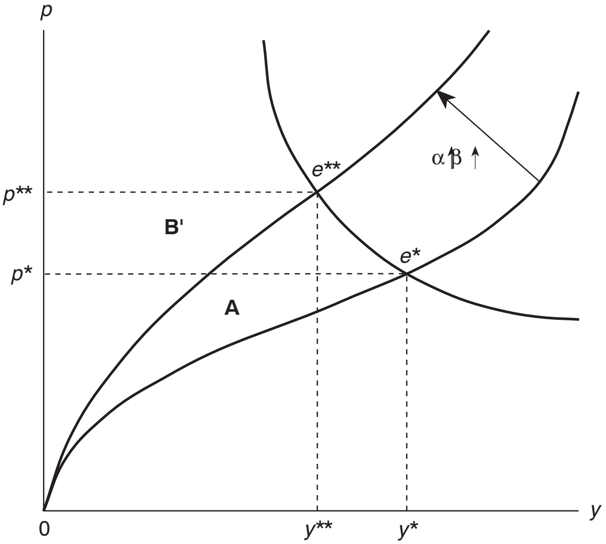


*Figure 7.A3* The general case

increase of cost gaps, and $B'$ represents the 'gain' due to an increase in the long-run equilibrium price gap. However, $B'$ in Figure 7.A3 is not as large as $B$ in Figure 7.A2, for the price elasticity of the demand allows the effect of cost increases to be absorbed not only by price hike but also by quantity reduction. This means that when the demand curve is steeply sloped, the gain component $B'$ is likely to outweight the loss component $A$. But, when the demand curve becomes more elastic, $B'$ becomes smaller, and in the limiting case of perfectly elastic demand curve it shrinks to zero.

This graphical explanation can be formalized as follows. First write down the relative form of industry demand function as $y_t = \tilde{D}(p_t)$. Then, $p^*$ is determined by the supply-demand equation: $\tilde{S}(p^*) = \tilde{D}(p^*)$. Differentiating this with respect to $\alpha$ and $\beta$ and rearranging terms, we have: $\partial p^*/\partial\alpha = (-p^*\partial\tilde{S}(p^*)/\partial\alpha)/(\varepsilon + \eta)$ and $\partial p^*/\partial\beta = (-p^*\partial\tilde{S}(p^*)/\partial\beta)/(\varepsilon + \eta)$, where $\varepsilon$ and $\eta$ are the price-elasticity of the supply curve and of the demand curve, respectively defined by $(\partial\tilde{S}(p)/\partial p)/(\tilde{S}(p)/p)$ and $(\partial\tilde{D}(p)/\partial p)/(\tilde{D}(p)/p)$. Keeping this in mind and differentiating (24), we obtain:

$$\begin{aligned} \frac{\partial r^*}{\partial\alpha}\bigg|_{\tilde{S}(p^*)=\tilde{D}(p^*)} &= b\int_0^{\tilde{S}^{-1}(y^*)} \frac{\partial\tilde{S}(z)}{\partial\alpha}\mathrm{d}z + bp^*\frac{-\partial\tilde{S}(p^*)}{\partial\alpha}\frac{1}{\eta+\varepsilon} \equiv -A_\alpha + B'_\alpha.; \\ \frac{\partial r^*}{\partial\beta}\bigg|_{\tilde{S}(p^*)=\tilde{D}(p^*)} &= b\int_0^{\tilde{S}^{-1}(y^*)} \frac{\partial\tilde{S}(z)}{\partial\beta}\mathrm{d}z + bp^*\frac{-\partial\tilde{S}(p^*)}{\partial\beta}\frac{1}{\eta+\varepsilon} \equiv -A_\beta + B'_\beta. \end{aligned} \tag{7.A3}$$

Note that the component $B'$ in either expression is decreasing in $\eta$. In particular, when $\eta = \infty, B'$ becomes equal to 0; when $\eta = 0, B'$ becomes equal to $B$ in (7.A2). Hence, we have obtained (7.31) of the main text.

## Acknowledgement

This is a simplified version of the paper I presented at ISER XI Workshop at the Certosa di Pontignano, Siena, on July 1, 1998. Since the workshop paper developed a series of new evolutionary models, it was very long and mathematically complicated. The present work has used the simpler evolutionary model of Iwai (1984b) so that I can present the same thesis much more economically. A revised version of the workshop paper is to be published as 'A Contribution to the Evolutionary Theory of Innovation, Imitation and Growth' in *Journal of Economic Behavior and Organization*. I am grateful to the participants of the Siena Workshop for their suggestions as well as to the members of Macro Workshop at University of Tokyo for their comments. The remaining errors are exclusively mine.

## Notes

1 Schumpeter (1939), p.100.

2 Schumpeter (1950), p. 83.

3 Schumpeter (1950), p. 28.

4 See, for instance, Nelson and Winter (1982), Dossi, Freeman, Nelson, Silverberg and Soete (1988), Metcalfe and Saviotti (1991), and Anderson (1994) for the comprehensive expositions of the 'evolutionary perspective' in economics.

5 The term 'satisficing' was first coined by Simon (1957) to designate the behavior of a decision maker who does not care to optimize but simply wants to obtain a satisfactory utility or return. The notion of 'organizational routines' owes to Nelson and Winter (1982). Organizations 'know' how to do things. In Iwai (1999) I have provided a legal-economic-sociological framework for understanding the nature and sources of such organizational capabilities.

6 Or we can think of this as a one-commodity economy with many competing firms.

7 Our evolutionary model can also accommodate a wide variety of industry structures. See Appendix A of Iwai (1984b) for the way to deal with the case of monopolistically competitive industry.

8 It is easy to show from (7.8) below that: $\dot{K}_t/K_t = \gamma(\log P_t - \sum_i (\log c_i) s_t(c_i)) - \gamma_0$, so that the growth rate of the industry's total capital stock is linearly dependent on the proportional gap between the price-wage ratio $P_t$ and the industry-wide average unit cost. If $\dot{K}_t/K_t$ is pre-determined (probably by the growth rate of the demand for this industry's products), this equation can be used to determine $P_t$. If, on the other hand, $P_t$ is pre-determined (probably by the labor market conditions in the economy as a whole), this equation can be used to determine $\dot{K}_t/K_t$. In either case, the forces governing the motion of $K_t$ are in general of the different nature from those governing the evolution of $\{S_t(c)\}$.

9 Schumpeter (1961), p. 154.

10 It is, however, not so difficult to deduce an investment function of this form by explicitly setting up an intertemporal optimization problem with adjustment costs, as in Uzawa (1969).

11 The actual derivation is as follows.

$$\begin{aligned} \dot{S}_t(c_i) \quad &\equiv \quad \sum_{j=i}^{n} \dot{s}_t(c_j) \\ &= \quad \sum_{j=i}^{n} (\dot{k}_t(c_j)/k_t(c_j) - K_t/K_t) s_t(c_j) \end{aligned}$$

$$\begin{aligned}
&= \sum_{j=i}^{n}\left((\gamma(\log p_t-\log c_j)-\gamma_0)-\sum_{h=1}^{n}(\gamma(\log p_t-\log c_h)-\gamma_0)s_t(c_h)\right)s_t(c_j) \text{ by (7.5)}\\
&= \sum_{j=i}^{n}\gamma\left(\sum_{h=1}^{n}(\log c_h)s_t(c_h)-\log c_j\right)s_t(c_j)\\
&= \gamma\left(\sum_{h=1}^{i-1}(\log c_h)s_t(c_h)S_t(c_i)-\sum_{h=i}^{n}(\log c_h)s_t(c_h)\,(1-S_t(c_i))\right)\\
&= \gamma\delta_t(c_i)S_t(c_i)(1-S_t(c_i)).
\end{aligned}$$

12 This is the simplification I also adopted in Iwai (1984b). However, in a recent article Franke (1998) indicated that the value of $\delta_t(c)$ may actually vary considerably as the parameter values of $\gamma, \nu$ and $\lambda$ as well as the value of $c$ vary. A caution is thus needed to use this approximation for purposes other than heuristic device.

13 A logistic differential equation: $x' = ax(1-x)$ can be solved as follows. Rewrite it as: $x'/x-(1-x)'/(1-x)$ and integrate it with respect to $t$, we obtain: $\log(x)-\log(1-x) = \log(x_0)-\log(1-x_0)+at$, or $x/(1-x) = \mathrm{e}^{at}x_0/(1-x_0)$. This can be rewritten as: $x = 1/(1+(1/x_0-1)\mathrm{e}^{at})$, which is nothing but a logistic equation given by (7.9).

14 The reason I have designated this *Hypothesis* by (*IM′*) is to differentiate it from a slightly different hypothesis adopted in Iwai (1984a). Its *Hypothesis (IM)* assumes that the probability of imitating a better technology is proportional to the frequency (rather than their capital share) of the firms using it. On the other hand, Iwai (2000) has adopted yet another hypothesis which assumes that firms imitate only the best practice technology and the probability of its success is proportional to the frequency of the firms using it.

15 p. 615.

16 See, for instance, Mansfield, Schwartz and Wagner (1981), Gorts and Klepper (1982) and Metcalfe (1988).

17 It is, however, possible to incorporate a trade-off between the resources devoted to capital growth and the resources devoted to imitative activities into our model. For instance, the growth parameters $\gamma$ and/or $-\gamma_0$ in (7.5) can be made a decreasing function of the imitation coefficient $\mu$.

18 The actual derivation is as follows. The value of $S_t(c_i)$ increases whenever one of the firms with unit costs higher than $c_i$ succeeds in imitating one of the technologies with unit costs $c_i$ or lower. Indeed, because of the assumption of the disembodied nature of technology, it increases by the magnitude equal to the imitator's capacity share. Note that $S_t(c_i)$ is not affected by the imitation of any of the firms with unit costs $c_i$ or less, for it only effects an infra-marginal transfer of capacity share. Let $M_t(c_i)$ denote the number of firms with unit costs $c_i$ or lower. Since the average capacity share of the firms with unit costs higher than $c_i$ is $(1-S_t(c_i))/(M-M_t(c_i))$ and the probability of a successful imitation for each of those $M-M_y(c_i)$ firms is $\mu S_t(c_i)\mathrm{d}t$ during a small time interval $\mathrm{d}t$, we can calculate the expected increase in $S_t(c_i)$ during $\mathrm{d}t$ as $((1-S_t(c_i))/(M-M(c_i))(\mu S_t(c_i)\mathrm{d}t)(M-M_t(c_i)) = (\mu S_t(c_i)\mathrm{d}t)(1-S_t(c_i)))$. If the number of firms is sufficiently large, the law of large numbers allows us to use this expression as a good approximation of the actual rate of change in $S_t(c_i)$. Dividing this by $\mathrm{d}t$ and letting $\mathrm{d}t \to 0$, we obtain (7.11).

19 Schumpeter (1961), p. 88.

20 Iwai (2000), however, presents an evolutionary model which does not separate innovators from inventors and assume that each innovation raises the productivity of the industry's best technology by a fixed proportion.

21 See, for instance, Kamien and Schwarts (1982), Grilliches (1984) and Scherer and Ross (1990).

22 It is, however, possible to incorporate a trade-off between the resources devoted to capital growth and the resources devoted to innovative activities into our model. For instance, the growth parameters $\gamma$ and/or $-\gamma_0$ in (7.5) can be made a decreasing function of the innovation coefficient $\nu$.

23 Iwai (1984a, 2000) also develops versions of evolutionary models which assume that only the firms currently using the best technology can strike the next innovation. In this case, the process of technological innovations is no longer a Poisson process, so that it is necessary to invoke the so-called 'renewal theory' in mathematical probability theory to analyze the long-run performance of the state of technology.

24 The derivation is as follows. Whenever one of the firms with unit costs higher than $c$ succeeds in innovation, the value of $S_t(c)$ increases by the magnitude equal to the innovator's capacity share. ($S_t(c)$ is, however, not affected by the innovation of any of the firms with unit costs c or less, because it only effects an infra-marginal transfer of the capacity share.) As in *note 11*, let $M - M_t(c_i)$ denote the total number of firms with unit costs higher than $c_i$. The average capacity share of the firms with unit costs higher than $c_i$ is $(1 - S_t(c_i))/(M - M_t(c_i))$ and the probability of a successful innovation for each of those $M - M_t(c_i)$ firms is $\nu \mathrm{d}t$ during a small time interval $\mathrm{d}t$. We can then calculate the expected increase in $S_t(c_i)$ due to an innovation as $((1 - S_t(c_i))/(M - M(c_i)))(\nu \mathrm{d}t)(M - M_t(c_i)) = (\nu \mathrm{d}t)(1 - S_t(c_i))$. If we divide this by $\mathrm{d}t$ and add to it the effects of economic selection and technological imitations given by (7.13), we obtain (7.18).

25 In deducing (7.19), we have employed a boundary condition$\hat{S}_{T(c)}(c) = \nu$ or $\hat{S}_{T(c)}(c) = 0$.

26 In our companion paper [1998] which assumes the step-by-step nature of innovations, it is $\beta$ that is assumed to be exogenously given.

27 The derivation is as follows.

$$\begin{aligned}\partial \tilde{S}(z)/\partial \alpha &= (1 - (1+\alpha)z/(\alpha\beta) - e^{-(1+\alpha)z/(\alpha\beta)})\alpha^{-2} e^{-(1+\alpha)z/\alpha\beta} \\ &\quad (1 + e^{-(1+\alpha)z/\alpha\beta}/\alpha)^{-2} < 0\end{aligned}$$

and

$$\partial \tilde{S}(z)/\partial \beta = -(1+\alpha)z/(\alpha\beta)^{-2} e^{-(1+\alpha)z/\alpha\beta}(1 + e^{-(1+\alpha)z/\alpha\beta}/\alpha)^{-2} < 0.$$

Note that $\nu$ appears both in $\alpha$ and in $\beta$. But its impact on $\tilde{S}(z)$ can be calculated as

$$\begin{aligned}\partial \tilde{S}(z)/\partial \nu &= \alpha \partial \tilde{S}(z)/\partial \alpha - \beta \partial \tilde{S}(z)/\partial \beta \\ &= 1/\alpha^{-1}(1 - e^{-(1+\alpha)z/\alpha\beta})e^{-(1+\alpha)z/\alpha\beta}(1 + e^{-(1+\alpha)z/\alpha\beta}/\alpha)^{-2} > 0.\end{aligned}$$

28 The derivation is as follows.

$$\begin{aligned}r^* &= b\int_0^{p^*} \tilde{S}(z)\mathrm{d}z \\ &= b\int_0^{\tilde{S}(p^*)} \tilde{S}(\mathrm{d}z/\mathrm{d}\tilde{S})\mathrm{d}\tilde{S} \\ &= b\int_0^{\tilde{S}(p^*)} \tilde{S}(\alpha\beta/((1-\tilde{S})(\alpha+\tilde{S})))\mathrm{d}\tilde{S} \\ &= (b\alpha\beta/(1+\alpha))\int_0^{\tilde{S}(p^*)} (1/(1-\tilde{S}) - \alpha/(\alpha+\tilde{S}))\mathrm{d}\tilde{S}\end{aligned}$$

Noting that $y^* = \tilde{S}(p^*)$, an explicit integration leads to (7.26).

29 In this sense, this section follows up the simulation exercises of Nelson and Winter in (1974) and chapter 9 of (1982).

30 See Sato (1975) for the general discussions on the aggregation of micro production functions.

31 The derivation of this *Proposition* is as follows. Since the short-run '*pseudo*' production function (7.27) implies that $l = \int_0^P c\mathrm{d}S_t(c)$ whenever $y = S_t(P)$, we have $\hat{l} = \int_0^P c\mathrm{d}\hat{S}_t(c)$ whenever $\hat{y} = \hat{S}_t(p)$. But from (7.21) we then have

$$\begin{aligned}\hat{l} &= \int_0^P c\mathrm{d}\tilde{S}(z) \\ &= \int_0^p \mathrm{e}^{z+\log C(t)}\mathrm{d}\tilde{S}(z) \\ &= \mathrm{e}^{-\lambda t}\int_0^p \mathrm{e}^{z}\mathrm{d}\tilde{S}(z) \\ &= \mathrm{e}^{-\lambda t}\int_0^{\tilde{S}(p)} \mathrm{e}^{z(\tilde{S})}\mathrm{d}\tilde{S} \\ &= \mathrm{e}^{-\lambda t}\int_0^{\tilde{S}(p)} ((\alpha+\tilde{s})/(\alpha(1-\tilde{S})))^{\alpha\beta/(1+a)}\mathrm{d}\tilde{S}\end{aligned}$$

and $\hat{y} = \hat{S}(p)$. Putting these two relations together, we obtain (7.29).

32 More precisely, we have $\mathrm{d}y/\mathrm{d}(l\mathrm{e}^{\lambda t}) = ((\alpha+y)/(\alpha(1-y)))^{-\alpha\beta/(1+\alpha)} > 0$ and $\mathrm{d}^2y/\mathrm{d}(l\mathrm{e}^{\lambda t})^2 = -\beta((\alpha+y)/(\alpha(1-y)))^{\alpha\beta/(1+\alpha)-1}/(1-y)^2 < 0$.

33 It is true that in the present model the rate of pure labor augmenting technical progress in the *pseudo*-aggregate production function is a given constant $\lambda$ which is determined exogenously by inventive activities outside of the industry. However, in some of the models presented in a companion paper (Iwai [2000]) this rate becomes also an amalgam of the parameters representing the forces of economic selection, technological diffusion and recurrent innovations.

34 This can be shown as follows. Let us differentiate (7.34) (or an equivalent expression given in *note* 33) with respect to $\alpha$. We then have: $0 = \mathrm{e}^{z(\tilde{S})}\partial\tilde{f}/\partial\alpha + \int_0^{\tilde{s}} \mathrm{e}^{z(\tilde{S})}(\partial z(\tilde{S})/\partial\alpha)\mathrm{d}\tilde{S}$. Since $z(\tilde{S})$ is an inverse function of $\tilde{S}(z)$ and $\partial\tilde{S}(z)/\partial\alpha < 0$ by (7.23), we have $\partial z(\tilde{S})/\partial\alpha > 0$. Hence we have $\partial\tilde{f}/\partial\alpha < 0$ as in (7.35). We can also show that $\partial\tilde{f}/\partial\beta < 0$ in exactly the same manner.

35 Schumpeter (1961), p. 154.

## References

Anderson, E.S. (1994) *Evolutionary Economics: Post-Schumpeterian Contributions*, Pinter: London.

Arrow, K. (1962) 'Economic welfare and the allocation of resources for inventions', in R. R. Nelson (ed.), *The Role and Direction of Inventive Activity*, Princeton University Press: Princeton.

Dossi, G., G. Freeman, R. Nelson, G. Silverberg and L. Soete (1988) (eds), *Technical Change and Economic Theory*, Pinter: London.

Franke, R. (1998) 'Wave trains and long waves: a reconsideration of Professor Iwai's Schumpeterian dynamics', forthcoming in D. Dell Gatti, M. Gallegati and A. Kirman (eds), *Market Structure, Aggregation and Heterogeneity*. Cambridge University Press: Cambridge.

Gorts, M. and S. Klepper (1982) 'Time-paths in the diffusion of product innovations', Economic Journal, 92, 630–53.

Grilliches, Z. (1984) (ed.) *R and D, Patents and Productivity*, Chicago University Press: Chicago.

Iwai, K. (1984a) 'Schumpeterian dynamics: an evolutionary model of innovation and imitation', *Journal of Economic Behavior and Organization*, 5, (2) 159–90.

Iwai, K. (1984b) 'Schumpeterian dynamics, Part II: technological progress, firm growth and "economic selection",' *Journal of Economic Behavior and Organization*, 5, (3) 321–51.

Iwai, K. (1999) 'Persons, things and corporations: Corporate personality controversy and comparative corporate governance', *American Journal of Comparative Law*, 45, 101–50.

Iwai, K. (2000) 'A contribution to the evolutionary theory of innovation, imitation and growth', *Journal of Economic Behavior and Organization*, 43, 167–98.

Kamien, M. and N. Schwarts (1982) *Market Structure and Innovation*, Cambridge University Press: Cambridge.

Maddison, A. (1987) 'Growth and slowdown in advanced capitalist economies: Techniques of quantitative assessment', *Journal of Economic Literature*, 25 (June 1987), 649–98.

Mansfield, E., M. Schwartz and S. Wagner (1981), 'Imitation costs and patents: an empirical study', *Economic Journal*, 91, (364), 907–18.

Metcalfe, J.S. 'The diffusion of innovation: an interpretive survey', Chapter 25 of G. Dossi *et al.* (1988).

Metcalfe, J.S. and P. Saviotti (1991) (eds) *Evolutionary Theories of Economics and Technological Changes*, Harwood Publishers: Reading. PA.

Nelson, R. and S. Winter (1974) 'Neoclassical vs. evolutionary theories of economic growth: critique and perspective', 84, 886–905.

Nelson, R. and S. Winter (1982) *An Evolutionary Theory of Economic Change*, Harvard University Press: Cambridge, MA.

Sato, K. (1975) *Production Functions and Aggregation*, North-Holland: Amsterdam.

Scherer, F.M. and D. Ross (1990) *Industrial Market Structure and Economic Performance*, Houghton and Mifflin: Boston.

Schumpeter, J.A. (1939) *Business Cycles*, McGraw-Hill: New York.

Schumpeter, J.A. (1950) *Capitalism, Socialism and Democracy*, 3rd. ed., Rand McNally: New York.

Schumpeter, J.A. (1961) *The Theory of Economic Development*, Oxford University Press: Oxford.

Simon, H.A. (1957) *Models of Man*, Wiley: New York.

Solow, R.M. (1956) 'A contribution of the theory of economic growth', *Quarterly Journal of Economics*, 70, (1), 65–94.

Solow, R.M. (1957) 'Technical change and the aggregate production function',*Review of Economics and Statistics*. 39, 312–20.

Uzawa, H. (1969) 'Time Preference and the Penrose Effect in a Two-Class Model of Economic Growth'. *Journal of Political Economy*, 77 (July/Aug.), 628–92.

# Part III

# Dynamics by interaction

# 8 Asymmetrical cycles and equilibrium selection in finitary evolutionary economic models

*Masanao Aoki*

## 8.1 Introduction

This chapter takes a fresh look at economic phenomena with multiple equilibria, and some of the issues associated with modeling asymmetrical business cycles. It summarizes an approach proposed in Aoki (1996, 1998), which is based on using continous time Markov chains and illustrates the advantages of using models with at most countable state spaces to answer these questions. The main advantage of employing models with a finite number of agents is simplicity in clarifying the issue of asymmetrical oscillation and naturalness of the criterion proposed for equilibrium selection. Two examples are presented to illustrate the proposed method. One is a binary choice model in which agents choose one of two alternative decisions or behavioral rules. Agents' decisions are interdependent due to externalities, such as congestion, bandwagon, group sentiments, network externalites and the like. The other example is a finite agent version of Diamond's search model, Diamond (1982).

## 8.2 Structure of transition rates

Dynamic behavior of a collection of (micro)economic agents is modeled by a continous time Markov chain, i.e., Markov process with at most countable states, called a jump Markov process. It describes time evolution of the probabilities of states of Markov chains by accounting for probability flows into and flows out of (sets of) states. The equations that do this accounting of probabilities are called the Chapman–Kolmogorov equations in the stochastic processes literature, see Whittle (1976, 156, 175), or Karlin and Taylor (1981, 286), for instance. We use a version of them which is easier to interpret. They are called the master equations in the physics and mathematical sociology literatures and we will use the same name in this chapter.[1] The master equations describe time evolution of probabilities of states of dynamic processes in terms of the probability transition rates and state occupancy probabilities. See Kelly (1979, 3) for example.

We must specify transition rates from a microeconomic consideration to describe the dynamics, and to draw macroeconomic implications of the Markov processes thus specified. With countable state spaces, transition rates specify well-behaved jump Markov processes, that is, processes do not execute an infinite number of jumps instantaneously. See Kendall (1975), Kelly (1976) or Breiman (1968).

In modeling a collection of interacting microeconomic units, henceforth called agents, the variables which are most important in influencing macroeconomic dynamics of the model are often not the absolute numbers of agents who occupy particular sets of states of the model, but rather the proportions or fractions of agents in these sets of states,[2] and possibly some function of the total number of agents, $N$, as an indicator of the scale or size of the model.

With state variable $X(t)$, the probability distribution is governed by the master equation

$$\frac{\mathrm{d}P(X,t)}{\mathrm{d}t} = \int \{w_N(X|X')P(X',t) - w_N(X'|X)P(X,t)\}\mathrm{d}X'$$

where we indicate $N$ explicitly as subscript to the transition rates from state $X$ to $X'$ as $w_N(X'|X)$. We give a heuristic derivation of this relation for discrete state variables.

When $X$ is discrete, the integral is replaced by summation. Let $X(t)$ be a state vector in a finite set $S$. Using the backward Chapman-Kolmogorov formulation we derive

$$\mathrm{d}P(X',t)/\mathrm{d}t = \sum_X P(X,t)w_N(X'|X,t),$$

which is known as the master equation. See Aoki (1996, 116) for a heuristic derivation. See Karlin and Taylor (1981, Chapter 14 ), for example, for the proof of the existence of the time derivative.

Write the sum of the right hand side separately for state $X'$ and the rest as

$$\sum_X P(X,t)w_N(X'|X,t) = \sum_{X\neq X'} P(X,t)w_N(X'|X,t) + P(X',t)w_N(X'|X',t),$$

and substitute $-\sum_{X\neq X'} w_N(X|X',t)$ in $w(X'|X',t)$ to rewrite the master equation as

$$\mathrm{d}P(X',t)/\mathrm{d}t = \sum_{X\neq X'} P(X,t)w_N(X'|X,t) - \sum_{X\neq X'} P(X',t)w_N(X|X',t).$$

This is the usual form in which the master equation is stated. The first term is the sum of the probability flows into state $X'$, and the second the probability flow out of state $X'$.

We assume that $x(t) = X(t)/N$ behaves as an intensive variable, and that the transition rates depend on $X$ and $N$ through $x$, except possibly for the scale factor which may depend on $N$.[3]

We make one additional key assumption that the change in the state variable, i.e. $X' - X$, remains the same for different values of $N$. This assumption is certainly met in birth-and-death processes, or birth-and-death with immigration processes, since jumps are restricted to be ±1 or some fixed integers from any state regardless of the total number $N$. This assumption actually is concerned with the scaling properties or homogeneity properties of the transition rates. Loosely put, it means that each of $N$ agents contribute approximately equally to the transition events. To make this explicit, express the transition rate as a function of the starting state, $X'$, and the jump (vector), $r = X - X'$ as

$$w_N(X|X') = w_N(X';X-X') = w_N(X';r),$$

and assume that

$$w_N(X';r) = w_N\{N(X'/N);r\} = Nw_N(X'/N;r) = N\Phi(x';r),$$

for some function $\Phi$, and where $x' = X'/N$. Using the same function we can express the transition rate in the opposite direction as

$$w_N(X'|X) = N\Phi(x;-r).$$

A scaling property which is seemingly more general is

$$w_N(X';r) = f(N)\Phi(x';r),$$

for some positive function $f(N)$. Actually, this factor $f(N)$ can be arbitrary, since it can always be absorbed into the choice of time unit.

More generally, the transition rates may take the form

$$w_N(X|X') = f(N)\{\Phi_0(x';r) + N^{-1}\Phi_1(x';r) + N^{-2}\Phi_2(x';r) + \cdots\}, \qquad (8.1)$$

where higher order terms in $N^{-1}$ may represent higher order interactions among microeconomic units beyond those which are captured by the leading term. In terms of these transition rates the master equation may be rewritten as

$$\frac{\mathrm{d}P(X,t)}{\mathrm{d}t} = f(N)\int\{\Phi_0(\frac{X-r}{N};r) + N^{-1}\Phi_1(\frac{X-r}{N};r) + \cdots)\}P(X-r,t)\mathrm{d}r$$
$$- f(N)\int\{\Phi_0(\frac{X}{N};r) + N^{-1}\Phi_1(\frac{X}{N};r) + \cdots\}P(X,t)\mathrm{d}r. \qquad (8.2)$$

Series expansions in terms of some fractional powers of $N$, such as $N^{-1/2}$ may also be used in some problems.

**Example 1: A binary choice model** Suppose that the total number of agents, $N$, is constant and that each of $N$ agents face a binary choice, $c_1$, and $c_2$. Merits or benefits of choices depend on the fraction of agents with particular choices. Let $n$ be the number of agents with $c_1$ at a point in time. This, or its fractional form $n/N$ serves as state variable. We suppress time argument for simplicity. If an agent can make his choice independent of others, then the transition rate from $n$ to

$n+1$ occurs because one of the agents with choice 2 has changed its mind, while $n$ becomes $n-1$ if one of the agents with choice 1 changes its mind. We assume no entry and no exit to keep $N$ fixed in this example, for simplicity. To account for the externality of decisions we posit

$$r_n := w_N(n+1|n) = a(N-n)\eta\left(\frac{n}{N}\right) = aN(1-x)\eta(x),$$

and

$$l_n := w_N(n-1|n) = bn\left[1-\eta\left(\frac{n}{N}\right)\right] = bNx[1-\eta(x)].$$

Without the factor $\eta$, each of $N-n$ agents changes its mind at the rate $a\Delta t$ over a small time interval $\Delta t$, and each of $n$ agents does likewise at the rate $b\Delta$. Here $a$ and $b$ are some constant. To simplify presentation we let them be the same.

The master equation is given, in terms of $n(t)$

$$\mathrm{d}P(n(t))/\mathrm{d}t = r_{n-1}P(n(t)-1) + l_{n+1}P(n(t)+1) - [r_n + l_n]P(n(t)),$$

together with boundary conditions at 0 and $N$, where $n(t)$ is the number of agents with the first choice. Equivalently it can be written in terms of $x(t) = n(t)/N$.

**Example 2: A finitary search model** To reformulate the search model of Diamond as a finitary evolutionary model, let $n$ be the number of employed. This, or its fractional form $e(t) = n(t)/N$ serves as state variable. The number increases from $n$ to $n+1$ when one of the unemployed encounter the production opportunity with production cost less than his reservation cost. The probability of this implies that

$$r_n := w_N(n+1|n) = (N-n)aG(c^*\left(\frac{n}{N}\right) = N(1-e)aG(c^*(e)),$$

where $e = n/N$ is the fraction of employed, using Diamond's notation, and $c^*(e)$ is the reservation cost when the fraction is $e$. See Diamond (1982) or Aoki (1999).

The number n decreases by two when two employed randomly encounter to exchange their output. This probability involves a chosen agent finding his trading partner as well as a possibility of two of the other employed forming a pair without his finding his partner. We define $b(e)$ so that

$$l_n := w_N(n-2|n) = \frac{N}{2}eb(e).$$

The master equation is

$$\mathrm{d}P(n(t))/\mathrm{d}t = r_{n-1}P(n(t)-1) + l_{n+2}P(n(t)+2) - (r_n + l_n)P(n(t)),$$

together with some boundary conditions. See Aoki and Shirai (2000) for exact expressions.

## 8.3 Two types of state variables

In these two examples the number of agents of some type ( with choice $c_1$ or employed) serves as state variable. In situations with $K$ types of agents, $K \geq 3$, we need a vector $\mathbf{n} = (n_1, \ldots, n_K)$ as state vector in general, where $n_i$ is the number of agents of type $i$ or choice $i$. More correctly, the vector with fractions $n_i/N$ as the $i$th component, $i = 1, \ldots, K$ is a state vector. When the total number of agents is fixed, then a $K - 1$ dimensional vector may be used.

Although this choice of state variables seems quite natural, many situations can be described better using an alternative choice of state variables. This alternative choice is in line with the occupancy problems in statistics or in physics. See Feller (1970) for example. In this definition, think of types as boxes and agents as indistinguishable balls. That is, agents are treated as exchangeable, the labels attached to agents being non-intrinsic and for mere convenience. The state vector is defined by a vector $\mathbf{a}$ with components $a_i$ being the number of types (boxes) with exactly $i$ agents in each of them, $i = 1, \ldots, K$. By definition, then we have $\sum_i ia_i = N$, and $\sum_i a_i \leq K$. This type of state variables are useful in dealing with distributions of firms by size, for example. Here interpret size broadly such as the number of employees, amount of capital or output in some convenient units, and so on. We do not have space to discuss models of this kind of state variable description. See Aoki (1999, 2000) for examples.

## 8.4 Dynamics

We use dynamic equations for the probability distributions of the states as the basic dynamic description of economic models with many agents of several types or agents with discrete choices. We do not use ordinary differential equations for the states themselves. This distinction is important, if subtle. That is, we do not derive differential equations for $X(t)/N$, such as the fraction of the employed in Diamond (1982), for example, but rather the differential equation for the probability of the fraction of the employed. This is what we call the master equation. In Weidlich and Haag (1983) we find examples of a birth-and-death stochastic process, which has a simple master equation, being adapted to model diffusion of opinion or information such as brand choices of consumer goods or voting for political candidates among a population. Recent examples in which master equations are used in economic models are Kirman (1993), Weidlich (1994), and Aoki (1995, 1998, and 1999).

### *8.4.1 Equilibrium distribution*

We first describe models with scalar state variables. By setting the left hand side of the master equation to zero, we obtain the condition for a stationary solution. Let $P_e(x)$ denote a stationary probability distribution. In stationary state or in equilibrium the probability in- and out-flows balance at every state $y$, the relation

$$\sum_{x \neq y} w(y|x) P_e(x) = \sum_{x \neq y} w(x|y) P_e(y)$$

holds for all $y$. This is the balance condition of probability flows, called the full balance equation, Kelly (1979, 5).

If the probability flows balance for every pair of states, then the equation

$$w(y|x)P_e(x) = w(x|y)P_e(y)$$

holds for all $x$ and $y$. This is called the detailed balance conditions. See Kelly (1979, Sec.1.5) for Kolmogorov criteria for stationary Markov chains and processes to satisfy the detailed balance condition.

Given an irreducible Markov chain, for any state $x_i$ there is a finite sequence of states which reaches it from some initial state, $x_0, x_1, \ldots, x_i$. If the detailed balance condition holds, we have

$$P_e(x_i) = P_e(x_0) \prod_{k=0}^{i-1} [w(x_{k+1}|x_k)/w(x_k|x_{k+1})].$$

This probability distribution is a Gibbs distribution since we can express this as an exponential distribution

$$P_e(x) = \text{ constant } \exp\{-\beta N U(x)\},$$

with

$$-\beta N[U(x_i) - U(x_0)] = \sum \ln \left\{ \frac{w(x_{k+1}|x_k)}{w(x_k|x_{k+1})} \right\}, \tag{8.3}$$

where $\beta$ is a parameter introduced to embody uncertainty or imprecision involved in making decisions, see Aoki (1996, 138), for example. Note that (8.3) is independent of paths from $x_0$ to $x_i$, i.e. $U(x)$ is a potential. See the Kolmogorov criterion in Kelly (1979, 21).

**A binary choice model continued** Example 1 has the equilibrium distribution which can be put as

$$P_e(x) = B \exp[-\beta N U(x)],$$

where

$$-\beta N U\left(\frac{n}{N}\right) = \text{ const } + \sum_{k=1}^{n} \beta g\left(\frac{k}{N}\right) + N \ln C_{N,n} + O(1/N),$$

with

$$g\left(\frac{k}{N}\right) := \ln \frac{\eta\left(\frac{k-1}{N}\right)}{1 - \eta\left(\frac{k}{N}\right)} + O(1/N).$$

Parameter $\beta$ is introduced to indicate the degree of uncertainty about the relative merits of alternative choices. See Aoki (1996, 138) or Aoki (1998, 2000) for further discussion on this parameter.

We can show that the local minima of the potential corresponds to locally stable equilibria of the aggregate dynamic equation which is derived next.

In models with several types of agents or several choices, suppose we use $\mathbf{x}(t) = (x_1(t), \ldots, x_K(t))$ as the vector. The equilibrium distribution $P_e(\mathbf{x})$ in some cases has product form $\prod_k \pi_k(x_k)$. See Kelly (1979) and Pollett (1986) for several examples of this representation. Even when this product form is not exact, it may serve as an approximation in some cases. In terms of the alternative state vector **a** mentioned at the end of the previous section, there is multi-variate Ewens distribution which seems to hold much promise in economic modeling. Aoki (1996, 238) has some preliminary account of this distribution. In problems with agents of many types in which random partitions of agents into types or choices are important, the Ewens distributions invariably arise. Kingman (1978) defines the notion of random partitions. See Aoki (1999, 2000) for economic examples.

## 8.5 Approximate solutions of master equations

Only a special class of master equations admits closed form analytic solutions. Notable among them are generalizations of birth-and-death processes which are discrete state Markov chains in which states are integer-valued and jumps are restricted to be $\pm 1$.

When we cannot solve the master equation explicitly, we may approximately solve it by expanding the solution in some parameter such as the size of the model, $N$, in the transition rate expressions. The parameter should be such that it governs or influences the size of fluctuations of the probabilities by affecting the jumps. As the parameter value approaches a limit the size of fluctuations should approach zero, so that this solution method produces a macroeconomic (aggregative) equation which is appropriate for a model with a large number of agents.

We mention two methods for approximately solving the master equation, and use one to derive the macroeconomic (aggregate) dynamic equations. This method is in the time domain, utilizing the Taylor series expansion mentioned above. The other is in the transform domain, and uses probability generating functions and possibly their approximate solutions to derive differential equations for the first few moments, such as mean and variance via the cumulant generating functions. See Cox and Miller (1967, 158). We do not discuss this method here.

### *8.5.1 Power series expansion*

We base this subsection on van Kampen ( 1992). The idea is to expand the master equations in $N^{-1/2}$ retaining terms only up to the order $O(N^{-1})$.[4]

When we anticipate that the probability of the state will show a well-defined peak at some $X$ of the order $N$ and spread of the order $\sqrt{N}$, if the initial condition is

$$P(X,0) = \delta(X - X_0).$$

In such cases we change variable by introducing two variables $\phi$ and $\xi$, both of order one, and set (recall that $x(t) = X(t)/N$)

$$x(t) = \phi(t) + N^{-1/2}\xi(t).$$

If this change of variable decouples the equation for $\phi$ and that for $\xi$, and if the coefficients in the dynamics for $\xi$ involve only $\phi$, then we are in a position to derive a deterministic aggregate dynamic equation and a stochastic dynamic equation for the fluctuations around the aggregate equations. Later we show that $\phi$ is the mean of the distribution when this change of variable is applicable i.e. $\phi(t)$ keeps track of the mean of $x(t)$.

We next show that the scaling by the square root of $N$ introduced above is the right one because terms generated in the power series expansion of the master equation separate into two parts. The first part, which is the largest in magnitude, is an ordinary differential equation for $\phi$. This is interpreted to be the macroeconomic or aggregate equation. The remaining part is a partial differential equation for $\xi$ with coefficients which are functions of $\phi$, the first term of which is known as the Fokker–Planck equation.

To obtain the solution of the master equation, we may set the initial condition by $\phi(0) = X_0/N$.[5]

We rewrite the probability density for $\xi$ as

$$\Pi\{\xi(t),t\} = P\{X(t),t\},$$

by substituting $N\phi + N^{1/2}\xi$ into $X$.[6] In rewriting the master equation for $\Pi$ we must take the partial derivative with respect to time by keeping $x(t)$ fixed, i.e. we must impose the relation

$$\frac{\mathrm{d}\xi}{\mathrm{d}t} = -N^{1/2}\frac{\mathrm{d}\phi}{\mathrm{d}t},$$

and we obtain

$$\frac{\mathrm{d}P}{\mathrm{d}t} = \frac{\mathrm{d}\Pi}{\mathrm{d}t} = \frac{\partial\Pi}{\partial\xi}\frac{\mathrm{d}\xi}{\mathrm{d}t} = \frac{\partial\Pi}{\partial t} - N^{1/2}\frac{\mathrm{d}\phi}{\mathrm{d}t}\frac{\partial\Pi}{\partial\xi}.$$

We also note that we need to rescale time by

$$\tau = N^{-1}f(N)t.$$

Otherwise, the random variable $\xi$ will not be of the order $O(N^0)$ contrary to our assumption, and the power series expansion will not be valid. But, $f(N) = N$ in this section. In general $\tau \neq t$. We use $\tau$ from now on to accommodate this more general scaling function.

The master equation in the new notation is given by

$$\begin{aligned}\frac{\partial\Pi(\xi,\tau)}{\partial\tau} - N^{1/2}\frac{\mathrm{d}\phi}{\mathrm{d}\tau}\frac{\partial\Pi}{\partial\xi} \quad = \quad & -N^{1/2}\frac{\partial}{\partial\xi}\{\alpha_{1,0}(x)\cdot\Pi\} + \frac{1}{2}\frac{\partial^2}{\partial\xi^2}\{\alpha_{2,0}(x)\cdot\Pi\} \\ & -\frac{1}{3!}N^{-1/2}\frac{\partial^3}{\partial\xi^3}\left\{\alpha_{3,0}(x)\cdot\Pi - N^{-1/2}\frac{\partial}{\partial\xi}\alpha_{1,1}(x)\cdot\Pi\right\} \\ & + O(N^{-1}),\end{aligned}$$

where $x = \phi(\tau) + N^{-1/2}\xi$, and where we define the moments of the transition rates by

$$\alpha_{\mu,\nu} = \int r^{\mu}\Phi_{\nu}(x;r)dr. \tag{8.4}$$

See van Kampen (1992, 253) for the terms not shown here.[7]

In this expression we note that the dominant term $O(N^{1/2})$ on the both sides are equated. In the next section, the first moment of the leading term of the transition rate $w_N(X|X')$, $\alpha_{10}(\phi)$, will be shown to determine the macroeconomic equation.

## 8.6 Macroeconomic equation

In the power series expansion we collect and equate the largest terms on both sides. This produces

$$\frac{d\phi}{d\tau} = g(\phi), \tag{8.5}$$

where we rename $\alpha_{1,0}(\phi)$ as $g(\phi)$ for short, where $\alpha_{1,0}$, defined in (8.4), is the first moment of the function $\Phi_0$, which appears in the transition rate expression $w_N(X|X') = f(N)\Phi_0(x';r)$, with respect to $r$. This is a deterministic aggregate equation for the average, $X/N$, which is the limiting dynamics as the number of agents goes to infinity.

The zeros of the right-hand side of this function are the equilibria of the macroeconomic model.

**A binary choice model continued** With the transition rates of Example 1, the macroeconomic equation becomes

$$\frac{d\phi}{dt} = (1-\phi)\eta(\phi) - \phi[1-\eta(\phi)]. \tag{8.6}$$

In terms of $g(\cdot)$ introduced in Example 1, the critical points of the macroeconomic equation is

$$\exp[\beta g(\phi)] = \frac{\phi}{1-\phi}.$$

**A finitary search model continued** In Example 2, the macroeconomic equation becomes

$$\frac{d\phi}{dt} = (1-\phi)aG(c^*(\phi)) - \phi b(\phi). \tag{8.7}$$

This is the same equation as that in Diamond (1982), because this term represent the first term in the Taylor series expansion. The difference lies in the equation for the fluctuation which is absent in Diamond since he assumes $N$ to be infinite from the beginning.

One should realize that the method presented above applies to models with discrete state space. See Aoki (1995, 1998, and 1999) for examples.

### *8.6.1 Multiple equilibria*

The zeros of the right-hand side of the aggregate equation (8.5) are critical points and potential candidates for the equilibria. If the sign of the first derivative $g'(\cdot)$ is negative at a critical point, then that critical point is a locally stable equilibrium. If $g(\cdot)$ is continous, then multiple zeros are alternatingly locally stable and unstable. In the example in Aoki and Shirai (2000), $g(\cdot)$ is discontinous, and $g'$ is negative except at the point of discontinuity. This example is the case of two locally stable equilibria. That is, the dynamics have two basins of attractions, and the centers are locally stable. We return to this example later in connection with the problem of equilibrium selection.

After the equation for $\phi$ is determined, the remainder of the master equation governs the density of $\xi$. This equation is called the Fokker–Planck equation when terms of the order $O(N^{-1/2})$ and smaller are neglected.

Later we return to this equation and calculate, among other things, the mean of $\xi$ and show that it remains at zero under certain conditions, i.e., that the mean of $x(t)$ is given by $\phi(t)$, and $\xi$ describes the spread about the mean, as we have claimed earlier. The variable $\phi$ may be thought of as the peak (the maximum likelihood estimate) of the distribution for $x$ in a single peaked distribution, while $\xi$ keeps track of the 'spread' of the distribution about the peak.

In some models $\alpha_{1,0}(\phi)$ or $c_1(\phi)$ is identically zero, then $\phi(\tau) = \phi(0)$ for all non-negative $\tau$, and a small deviation in $\phi(0)$ does not decay to zero. In this case we need to expand the master equation in $X/N$ , and redefine $\tau$ by $N^{-2}t$. Then we are lead to what is known as the diffusion approximation. We do not discuss this case here. See Aoki (1996, Chap.5 ), for example.

### *8.6.2 Macroeconomic relations: case of infinite number of agents*

When $N$ becomes very large only the first terms in the Taylor series expansions in the right-hand side of (8.1) and (8.2) remain. Traditional macroeconomic relations are recovered in the limit of $N$ going to infinity. For example, (8.6) and (8.7) are the macroeconomic dynamic equations in terms of fractions of agents with infinity of agents assumed. For example, (8.7) is exactly the equation derived by Diamond.

## 8.7 Dynamics for fluctuations

After the terms which determine the macroeconomic equation are removed, and when we retain terms up to $O(N^{-1/2})$, then we are left with

$$\frac{\partial \Pi(\xi,t)}{\partial t} = -\alpha_{10}'(\phi)\frac{\partial(\xi\Pi)}{\partial \xi} + \frac{1}{2}\alpha_{2,0}\frac{\partial^2 \Pi}{\partial \xi^2}.$$

This equation is linear in $\xi$, and is called a linear Fokker-Planck equation, Aoki (1996, 137, 158). This equation can be used to obtain the dynamics for the mean and variance of $\xi$, or as in the case of Example 1, the distribution function for $\xi$ can sometimes be determined.

**A finitary search model continued** The dynamic equation is given by

$$\partial\Pi/\partial t = A(\phi)\partial(\xi\Pi) + C(\phi)\partial^2\Pi/\partial\xi^2$$

up to $O(N^{-1/2})$, with $A(\phi) = -\Phi'(\phi)$, where $\Phi(\phi) = (1-\phi)aG(c^*) - \phi b(\phi)$, and $C(\phi) = (1/2)(1-\phi)aG(c^*) + \phi b(\phi)$. Near locally stable equilibrium the derivative $\Phi'(\phi)$ is negative, that is $A(\phi)$ is positive, and the stationary distribution of this equation turns out to be normal with zero mean and variance $C(\phi)/A(\phi)$.

## 8.8 Equilibrium selection

There is a large body of literature on the issue of how to select an equilibrium in models with multiple equilibria. With an infinite number of agents, the dynamics are deterministic. Hence the basins of attractions are fixed and once a state is in one of the basins, the state does not leave that basin, unless there are some exogenous devices such as sudden changes in expectations which force the state to jump from one basin to another.

With a finite number of agents in stochastic models such as those in this chapter, the state oscillate persistently among different basins of attractions and generate generally asymmetrical oscilaltions. See appendix in Aoki (1998) for a simple two state example to illustrate this point. See also Aoki and Shinai (2000).

In the framework employed in this chapter, the probability that the model stays in different basins can be used to show which basin of the model the state is likely to settle. Two examples will illustrate the method.

### *8.8.1 Case of two locally stable equilibria*

We approximately evaluate the mean transition times from one basin of attraction to the other, and calculate the equilibrium probabilities that the employed fraction of population stays in each of the two basins of attraction. Moreover, we show that our analysis provides the basis for equilibrium selection in the deterministic version of the model. Since the states off the equilibria approach them exponentially fast, only the events of disturbances which are large enough to bring the states from one basin of attraction to another will force the model to move from one equilibrium to another.

*Approximate analysis*

First we recognize that we need to calculate only the event from one of the equilibrium state to the boundary between two basins of attraction, $\psi$, which is introduced in the example above. The reason is the same one used by van Kampen (1992) as quoted in Aoki (1996, 151). The time needed for $\phi$ to reach its equilibrium value, $\phi_1$, or $\phi_2$ depending on the initial value, is much shorter than the time needed to go from one basin of attraction to the other.

A quick way to see this is to solve the deviational equation for $\phi$. To be definite, suppose that $\phi$ is in the domain of attraction to $\phi_1$ and let $x := \phi - \phi_1$. Then, it is governed by

$$\mathrm{d}x/\mathrm{d}\tau = \Phi_1'(\phi_1)x = -A(\phi_1)x,$$

with the initial condition $x(0) = \phi(0) - \phi_1$.

The solution is $x(\tau) = x(0)\exp[-A(\phi_1)\tau]$. Recalling that $\exp(-4.5) = 0.01$, it takes about $\tau = 4.5/A(\phi_1)$ to reduce the distance from $\phi_1$ to about 0.01 of the original value. In the case where $a = 1$ and $p = 0.2$, we have $\phi_1 = 0.358$ and $A(\phi_1) = 0.92$. Thus, it takes about 4.9 or 5 time units to reach the equilibrium point. As we show later in example 2, the mean first passage time for this example is of the order of $10^3$ when $N = 100$. Therefore we are justified in assuming that $\phi$ is initially at one of the equilibrium points in calculating the mean first passage time. The procedure is as outlined in Appendix in Aoki (1998). We set up a two state Markov chain, because there are two locally stable equilibria in the example. Let $\pi_1$ and $\pi_2 = 1 - \pi_1$ be the probability that the employed fraction are in basins of attraction for $\phi_1$ and $\phi_2$, respectively. These probabilities evolve according to the differential equation,

$$\mathrm{d}\pi_1/\mathrm{d}\tau = W_{2,1}\pi_2 - W_{1,2}\pi_1,$$

where $W_{1,2}$ is the transition rate from $\phi_1$ to the boundary of the two basins of attractions, that is, $\psi$, and $W_{2,1}$ is that from $\phi_2$ to $\psi$. In the stationary state $d\pi_1/d\tau = 0$, we have

$$\pi_1 = \frac{1}{W_{1,2}/W_{2,1} + 1}.$$

The mean first passage time is given by

$$\tau_{1,2} = \frac{1}{W_{1,2}}.$$

See Aoki (1996, 152).

To calculate $W_{1,2}$ we use the probability that

$$W_{1,2} = \Pr[\xi \geq \xi_c],$$

with

$$\phi_1 + \frac{\xi_c}{\sqrt{N}} = \psi.$$

Analogously $W_{2,1}$ is approximated by the probability that $\xi$ is smaller than $\sqrt{N}(\psi - \phi_2)$, or equivalently it is larger than $\sqrt{N}(\phi_2 - \psi)$.

### *8.8.2 Limiting case of N infinity*

As we increase the number of agents in the economy to infinity, our model converges to that of Diamond. This can be seen by the fact that variance of employed fraction of agents converges to zero as $N$ taken to infinity as it is indicated by the density function . This suggests that the stationary (invariant) distribution over fraction of employed agents in the economy become spikes with probability masses of $\pi_1$ and $\pi_2$ assigned for employed fractions $e = \phi_1$ and $e = \phi_2$, respectively.

These probability masses for each locally stable critical points provide the criteria for equilibrium selection for the model of multiple equilibria with infinite number of agents. One can easily check that our special case given in section 8.6 yields exactly the same stationary fractions of employed agents $\phi_1$ and $\phi_2$ in Diamond's model if we set the same matching function $b$ and cost distribution function $G$.

What we are left to do is to calculate $\pi_1$ when $N$ is taken to infinity. As suggested above, we have

$$\begin{aligned} W_{1,2} &= \Pr\left[\xi > \sqrt{N}(\psi - \phi_1)\right] \\ &= \int_{\sqrt{N}(\psi-\phi_1)}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2(\phi_1)}} \exp\left[-\frac{\xi^2}{2\sigma^2(\phi_1)}\right] \mathrm{d}\xi, \end{aligned}$$

and

$$\begin{aligned} W_{2,1} &= \Pr\left[\xi < \sqrt{N}(\psi - \phi_2)\right] \\ &= \int_{\sqrt{N}(\phi_2-\psi)}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2(\phi_2)}} \exp\left[-\frac{\xi^2}{2\sigma^2(\phi_2)}\right] \mathrm{d}\xi. \end{aligned}$$

It is easy to see that both $W_{1,2}$ and $W_{2,1}$ approach zero as $N$ is brought to infinity. Hence, we can approximate $\lim_{N\to\infty} W_{1,2}/W_{2,1}$ by $\lim_{N\to\infty}(dW_{1,2}/dN)/(dW_{2,1}/N)$. This is given by,

$$\begin{aligned} \lim_{N\to\infty} \frac{W_{1,2}}{W_{2,1}} &= \lim_{N\to\infty} \frac{\mathrm{d}W_{1,2}/\mathrm{d}N}{\mathrm{d}W_{2,1}/\mathrm{d}N} \\ &= \lim_{N\to\infty} \exp\left[\left(\frac{(\phi_2-\psi)^2}{2\sigma^2(\phi_2)} - \frac{(\psi-\phi_1)^2}{2\sigma^2(\phi_1)}\right)N\right] \frac{(\psi-\phi_1)\sigma(\phi_2)}{(\phi_2-\psi)\sigma(\phi_1)}. \end{aligned}$$

From this, it is straightforward to see that as $N$ approaches infinity, $W_{1,2}/W_{2,1}$ approaches 0 if and only if $(\psi-\phi_1)/\sigma(\phi_1) > (\phi_2-\psi)/\sigma(\phi_2)$; infinity if and only if $(\psi-\phi_1)/\sigma(\phi_1) < (\phi_2-\psi)/\sigma(\phi_2)$; and 1 if and only if $(\psi-\phi_1)/\sigma(\phi_1) = (\phi_2-\psi)/\sigma(\phi_2)$.

The result is summarized as follows;

$$\lim_{N\to\infty} \pi_1 = \begin{cases} 1 & \text{if } (\psi-\phi_1)/\sigma(\phi_1) > (\phi_2-\psi)/\sigma(\phi_2), \\ 1/2 & \text{if } (\psi-\phi_1)/\sigma(\phi_1) = (\phi_2-\psi)/\sigma(\phi_2), \text{ and} \\ 0 & \text{if } (\psi-\phi_1)/\sigma(\phi_1) < (\phi_2-\psi)/\sigma(\phi_2). \end{cases}$$

The larger the distance between the critical point and the boundary of basins of attraction, or the smaller the variance of fluctuation around the critical point, it is likely that this critical point is selected as an equilibrium in a model with infinite number of agents.

### *8.8.3 Case of two locally stable and one unstable equilibria*

We discuss the first passage between two locally stable equilibria when they are on both sides of a locally unstable equilibrium. The method for calculating first passage probabilities or mean first passage time is well known, and is discussed in Cox and Miller (1965, Section 3.4), Grimmett and Stirzaker (1992, Section 6.2), or van Kampen (1992, Chapter XII) to mention a few textbooks.

Basically, in going from the basin containing one locally stable equilibrium to the other basin, the model must overcome or go over the potential barrier $U_b - U_c$ or $U_b - U_a$, where we use the short-hand notation to denote $U(\phi_a)$ by $U_a$, and so on, where the three critical points are states,$\phi_a < \phi_b < \phi_c$, and $\phi_b$ is the unstable equilibrium point. That is, the model must go over the potential of height $U(b)$ in going from one basin to another. The model starts from potential height of $U(a)$ in one direction and $U(c)$ in another. Suppose that $U(a) < U(c)$. Since the equilibrium probability of going from $\phi_a$ to $\phi_c$ is proportional to $\exp[-\beta(U_b - U_a)]$, and the reverse direction has the probability proportional to $\exp[-\beta(U_b - U_c)]$, the model is more likely to stay in the basin containing state $\phi_a$ than that containing state $\phi_b$. Put differently, the transition from state a to state c takes longer than the reverse direction.

In the limit of $N$ going to infinity, the model will stay longer and longer in the basin containing state $\phi_a$, and in the limit this is the equilibrium selected.

To make this heuristic argument rigorous, we need to calculate the first passage times from one basin to another and its reverse and calcualte the probability that the model stays in each of the two basins. See Aoki (1996, Sec. 5.11)

## 8.9 Concluding remarks

We have proposed a probabilitic way for dealing with multiple equilibria and the associated asymmetrical cycles by formulating the dynamics for jump Markov processes as the master equation. Out of the master equation we derive deterministic aggregate or macroeconomic dynamics which determine locally stable equilibria as the centers of basins of attractions. The ratios of the probabilities of individual basins of attractions determine which of the equilibria is selected in the limit of the size of the model (the number of agents in the model ) goes to infinity.

In the limit of $N$ going to infinity, the aggregate dynamics become the same as those derived by assuming at the outset that there are infinity numbers of agents and working with appropriate fractions. You might ask, therefore, what is the advantage of the proposed procedure? The advantage lies in the way the model naturally introduces asymmetrical cycles, and selects one basin out of several as the one with most probability mass as $N$ is increased.

## Notes

1 There are many references on master equations. Van Kampen (1992, 97) tells us the origin of the word ‘master’. See also Weidlich and Haag (1983), or Kubo (1975).

2 This fraction is an example of variables called intensive variables. This is often implicit in the search literature, where it is routinely assumed that there are an infinite number of agents, and fractions of one kind or another are posited from the beginning, such as the fraction of employed, or fractions of agents which hold some specified assets, and so on. See Diamond (1982) or Kiyotaki and Wright (1993) for example.

3 When the size $N$ is fixed, this assumption is innocuous because it can be absorbed into time units. If $N$ is a random variable as in models open to entry and exit by agents, it must be explicitly incorporated into the transition rates. Kelly (1979) has an example in which $N$ is a random variable.

4 When these terms are zero, we may want to retain terms of the order $N^{-2}$. Then we have diffusion equation approximations to the master equation. Diffusion approximations are not discussed here.

5 We need not be precise about the initial condition since an expression of the order $O(N^{-1})$ or $O(N^{-1/2})$ can be shifted between the two terms without any consequence. Put differently, the location of the peak of the distribution can't be defined more precisely than the width of the distribution which is of the order $N^{1/2}$.

6 Following the common convention that the parameters of the density are not carried as arguments in the density expression, we do not explicitly show $\phi$ when the substitution is made.

7 Note that

$$\frac{\partial}{\partial\xi}\{\alpha_{10}(x)\Pi\} = \alpha_{10}(\phi)\frac{\partial\Pi}{\partial\xi} + N^{-1/2}\alpha'_{10}(\phi)\frac{\partial}{\partial\xi}(\xi\Pi) + O(N^{-1}).$$

## References

Aoki, M. (1995) 'Economic Fluctuations with Interactive Agents: Dynamic and Stochastic Externalities', *Japanese Economic Review*, **46**, 148–65.

Aoki, M. (1996) *New Approaches to Macroeconomic Modeling: Evolutionary Stochastic Dynamics, Multiple Equilibria, and Externalities as Field Effects* , Cambridge University Press, New York.

Aoki, M. (1998), 'Simple Model of Asymmetrical Business Cycles: Interactive Dynamics of a Large Number of Agents with Discrete Choices', *Macroeconomic Dynamics*, **2**, 427–42.

Aoki, M. (1999), 'Open Models of Share Markets with Several Types of Participants', presented at Workshop on Heterogenous Interacting Agents, Univ. Genoa, June.

aoki, M. (2000) 'Cluster Size Distributions of Economic Agents of Many Types in a Market', *Journal of Mathematical analysis and Applications*, **249**, 32–52.

Aoki, M. (2001) *Modeling Agregate Behavior and Fluctuations in Economics: Stochastic Views of Interacting Agents*, Cambridge University Press, Cambridge.

Aoki, M. and Y. Shirai (2000), 'Stochastic Business Cycles and Equilibrium Selection in Search Equilibrium', *Macroeconomic Dynamics*, 4, 487–505.

Breiman, L. (1968) *Probability Theory*, Addison-Wesley, Reading, MA.

Cox, D. R., and H. D. Miller (1967) *The Theory of Stochastic Processes*, Chapman & Hall, London.

Diamond, P. D. (1982) 'Aggregate Demand Management in Search Equilibrium'. *Journal of Political Economy*, **90**, 881–94.

Feller, W. (1970) *Introduction to Probability Theory and Its Applications, Vol I*, John Wiley & Sons, New York.

Grimmett, G. R. and D. R. Stirzaker (1992) *Probability and Random Processes*, Oxford University Press, New York.

Karlin, S. and H.Taylor (1981) *A Second Course in Stochastic Processes*, Academic Press, New York.

Kelly, F. (1979) *Reversibility and Stochastic Networks*, Wiley & Sons, New York, 1979

Kelly, F. (1976) 'On Stochastic Population Models in Genetics', *J. Appl. Probab.*, **13**,127–31.

Kendall, D. G. (1975) 'Some problems in mathematical genealogy', in J. Gani (ed.) *Perspective in Probability and Statistics*, Academic Press, New York.

Kingman, J. F. K. (1978) 'Random Partitions in Population Genetics', *Proc. London. Ser. A*, **361**, 1–20.

Kirman, A. (1993) 'Ants, Rationality, and Recruitment', *Quart.J.Econ.*, **103**, 137–56.

Kiyotaki, N., and R. Wright (1993) ' A Search-Theoretic Approach to Monetary Economics', *Amer. Econ. Review*, **83**, 63–97.

Kubo, R. (1975) 'Relaxation and Fluctuation of Marovariables' in Lecture Notes in Physics No.39, H. Araki (ed.), Springer-Verlag, Berlin.

Pollett, P. K. 1986) 'Connecting Reversible Markov Processes', *Adv. Appl. Probab.* **18**, 880–900.

van Kampen, N. G. (1965) 'Fluctuations in Nonlinear Systems', Chapter V in R.G. Gurgess (ed.) *Fluctuation Phenomena in Solids*, Academic Press, New York.

van Kampen, N. G. (1992) *Stochastic Processes in Physics and Chemistry: Revised and enlarged edition* North Holland, Amsterdam.

Weidlich, W. (1994) 'Synergetic Modelling Concepts for Sociodynamics with Application to Collective Political Opinion Formation', *J.Math.Sociology*, **18**, 267–91.

Weidlich, W. and G. Haag (1983) *Concepts and Models of a Quantitative Sociology*, Springer-Verlag, Berlin/Heidelberg/New York.

Whittle, P. (1976) *Probability via Expectation*, Second Edition, Springer-Verlag, New York.

# 9 The instability of markets

*Tad Hogg, Bernardo A. Huberman and Michael Youssefmir*

## 9.1 Introduction

The explosive growth of computer networks and of new forms of financial products, such as derivatives, is leading to increased couplings among previously dispersed markets. This increased fluidity raises questions about the stability and efficiency of the international financial system. On the one hand, it is quite apparent that increased overall connectivity among markets allows transactions that were not previously possible, increasing the net wealth of people and leading to more efficient markets. But at the same time, this implies that a transition is taking place, from a more static scenario in which isolated markets can be considered in equilibrium on their own, to a global economy that knows of no geographical borders. One may wonder about the nature of this transition, i.e. is it gradual in the sense of a smooth change in prices, or punctuated by abrupt changes in the value of certain commodities, cascading bankruptcies and market crashes?

Underlying these questions is the old problem of the existence and stability of equilibria in markets, which general equilibrium theory has addressed under various conditions (Arrow and Hahn 1971, Arrow 1988) . In this chapter, we will focus on the stability of markets by treating the couplings among them as dynamical entities in the spirit of other evolutionary approaches (Day 1975, Nelson and Winter 1982, Brock 1988) .

The standard approach to characterizing the stability of markets postulates that agents opportunistically optimize their portfolios in such a way so as to minimize their risk while maintaining a given return. But in real life agents are not always able to perfectly process information about the system in which they are embedded. Under such conditions, adaptive agents continuously switch between different behavioral modes in response to a constantly changing environment, favoring behaviors that can temporarily lead to increased rewards. It follows that while agents are continuously learning about the relative merits of different commodities and markets, the couplings between such commodities evolve in rather complicated ways.

Other examples of such evolving couplings are provided by agents linking limited baskets of commodities efficiently while ignoring other commodities in the process. Of more recent interest are the couplings that agents introduce through complicated derivatives created in an effort to optimize and hedge away risks. Due to their complex nature such derivatives can be poorly understood by the people who use them, once again leading to extraneous couplings. Finally, highly leveraged hedge funds can also introduce couplings into the market by being forced to cover certain leveraged positions with other unrelated positions.

In this chapter, we assume the existence of an equilibrium within a network of markets and examine its dynamic stability through the use of a general model of the adjustment processes within the markets. We focus on cases where agents are not perfectly rational or identical and therefore can introduce couplings that may not always represent the best possible allocation of their resources. We then discuss how the stability of such a system scales as different markets become increasingly coupled. We show that the class of systems that are stable becomes smaller and smaller as the number of coupled markets scales up. These instabilities are shown to exist even when couplings are so weak that the markets are near decomposable. Such instabilities in turn require a heightened degree of learning on the part of market participants, a process in which the learning itself is marked by periods of instability until equilibrium is reestablished. These results, which run counter to prevailing notions of stability in large coupled markets, offer a cautionary note on treating emerging markets with the tools of equilibrium economics.

## 9.2 Dynamics of coupled markets

In what follows we will consider the case of a number of markets, each of which is in stable equilibrium when decoupled from the others. These markets contain arbitrarily large number of agents that buy and sell commodities, using a diverse set of strategies. Heterogeneous agents will provide diverse couplings between these markets through mechanisms such as physical substitutions, financial derivatives, arbitraging between geographically disperse markets, and expectations (be they rational or irrational) that movements in one market will lead to changes in others.

Given a single commodity, a reasonable dynamic model of price adjustments postulates that the price rises when demand exceeds supply and that they fall when supply exceeds demand. This can be described by a differential equation of the form

$$\frac{\mathrm{d}p}{\mathrm{d}t} = f(p) \tag{9.1}$$

where $p$ is the price of the commodity and the function $f$ represents the excess demand at the given price. This equation can be thought of as a description of market adjustments via the *tatonnement* process by which an auctioneer calls out adjustments in prices in order to satisfy excess demand. The function $f$ then defines the equilibrium $p^{(0)}$ where $f(p^{(0)}) = 0$. The value $f'(p^{(0)})$ represents a linearization of the adjustment process near the equilibrium. If this quantity is negative the equilibrium at $p^{(0)}$ is stable against small perturbations, which relax on a time scale given by $\tau = \left|f'(p^{(0)})\right|^{-1}$.

The time scale $\tau$ for adjustment is reduced when transaction costs are negligible or agents are confident about the value of the commodity. Any movement of $p$

away from $p^{(0)}$ will be immediately corrected for by agents seeking to make a profit on the differential $p - p^{(0)}$. On the other hand, when agents are uncertain or have diversity of expectaions for the fair price for the commodity, the adjustment process may depend on agent interactions, communication, and analysis. In this case, the time scale for adjustment $\tau$ will tend to be greater than before.

These remarks are readily generalized to the adjustment processes where there is more than one commodity and multiple markets. We will study the dynamics that result from the couplings of these markets, and in particular how the couplings affect the stability of the system as whole. To do this we interpret $p$ as a price vector, whose entries correspond to prices in the separate markets, and $f$ as a vector of excess demands that characterizes both the dynamics of the isolated markets and their couplings. The dynamics is then given by an equation of the form (9.1), which relates the evolution of prices of all commodities in all markets as a function of each other and the degree to which they are coupled (Arrow and Hurwicz 1958, Samuelson 1941, Metzler 1945).

The excess demand in the case of each commodity is controlled by the individual agents within the respective markets. In stable equilibrium the individual excess demands of the agents are such that in totality prices are held at the equilibrium values. We can then study stability around an equilibrium $\vec{p_{\mathrm{eq}}}$ where $\vec{f}(\vec{p_{\mathrm{eq}}}) = 0$, by assuming the existence of a temporal departure from this equilibrium by a small amount $\vec{\delta}$. In order to see if this disturbance decays in time or grows, we perform a Taylor series expansion around the fixed point to get an equation of the form

$$\frac{\mathrm{d}\vec{\delta}}{\mathrm{d}t} = M(\vec{p_{\mathrm{eq}}})\vec{\delta} \tag{9.2}$$

where the $n$ by $n$ matrix $M$ is the Jacobian of $\vec{f}$ evaluated at the fixed point $\vec{p_{\mathrm{eq}}}$ and $n$ is the number of commodities.

The components of this Jacobian matrix describe how a small increase in the price of one item in one market changes that of another item. Thus, the diagonal elements of $M$ show the direct effect of a small change in excess demand on the price of a given commodity. We assume that each market by itself is stable so as to counteract the original change, which implies that the diagonal elements will be negative, with average value that we denote by $-D$, and which quantifies the speed of adjustments within that market.

On the other hand an off-diagonal entry describes the direct effect on an item from a change in the price of another, which can be of either sign. Notice that if the couplings between commodities are weak, it would translate into a matrix whose largest entries are on the diagonal and the off-diagonal terms would be small. As discussed above, these off-diagonal couplings are the result of individual agents whose expectations, whether informed or uninformed, link the two commodities such that an increase in one commodity's price affects that of the other.

The effect of slight disturbances around the fixed point perturbation is determined by the eigenvalue of $M$ with the largest real part, which we denote by $E$. Specifically, the long time behavior of the perturbation is given by $\delta \propto e^{Et}$, implying that if $E$ is negative, the disturbance will die away and the system will return to its original equilibrium. If, on the other hand, $E > 0$, the smallest perturbation will grow rapidly in time, leading to instability.

We note that this model of dynamic adjustment and its stability has also been used in the economic literature focusing on qualitatively specified matrices (Quirk 1981). In that literature, the focus is on specifying the stability properties of matrices given only the qualitative nature of the signs of the entries (positive, negative, or zero). In that case, for example, commodities that are gross substitutes would be coupled via positive nondiagonal matrix elements while gross compliments would be coupled by negative coefficients. Such an approach, however, suffers from the disadvantage that the cases for which stability criterion are specified are quite restricted.

## 9.3 Market instabilities

The approach taken in this chapter looks at the average behavior of a market given an ensemble specifying uncertain knowledge of parameters within the stability matrix. For the sake of treating a very general case, we will assume as little knowledge about these mechanisms as it is possible. This implies that all matrices that are possible Jacobians can be considered, and that there is no particular basis for choosing one over the other. This is the class of the so-called random matrices, in which all such matrices are equally likely to occur. Matrices in this class are such that each entry is obtained from a random distribution with a specified mean and variance. By taking this approach, we can make general statements about the stability of these systems as the number of markets and diversity of agent behaviors grows. We show quite generally that as the number of coupled markets and the diversity among the agents grows, it is more and more likely that the system as a whole will be unstable.

The precise value of the largest eigenvalue $E$ depends on the particular choice of the Jacobian matrix. Methodologically, the study of the general behavior of these matrices is performed by examining the average properties of the class that satisfies all the known information about them. A class of plausible stability matrices is determined by the amount of information one has about particular market mechanisms and their couplings. This information, which is far from perfect, depends on the nature of expectations that agents have about future values, which are based on limited knowledge, and on the extent to which individuals understand the nature of the financial instruments that they invest in.

In spite of their random nature, these matrices possess a number of well defined properties, among them the behavior of their eigenvalues (Wigner 1958, Mehta 1967, Cohen and Newman 1984, Furedi and Komlose 1981, Juhasz 1982, Edelman 1989). This means that we can use these properties to ascertain the stability of the markets against perturbations in the excess demand. In what follows we will show that in the general case, as these matrices grow in size, or the variability of their entries, their largest eigenvalue becomes positive, thus leading to market instability. This is a result that applies not only to markets but also to complex ecosystems (May 1972, McMurtie 1975, Hogg, Huberman and McGlade 1989).

The simplest case, albeit not very realistic, would correspond to a situation where commodities are equally likely to be substitutes or complements of each other. This means that on average the nondiagonal elements would be of zero mean. We will also assume that the Jacobian random matrix will have symmetric entries and be bounded in magnitude. For this case, the distribution of eigenvalues as a function of the size of the Jacobian is given by Wigner's law (Wigner 1958), i.e.,

$$E = 2\sigma\sqrt{n} - D \tag{9.3}$$

This implies that as the matrix gets large enough, its largest eigenvalue will become positive.

A more realistic case relaxes the requirement that the changes in one price are on average balanced by changes in others. This corresponds to having a non-zero mean. In this situation a theorem of Furedi and Komlos (1981) states that, on average, the largest eigenvalue is given by

$$E = (n-1)\mu + \sigma^2/\mu - D \tag{9.4}$$

where $\mu$ is the average value of the couplings, which we assume to be positive. Moreover, the actual values (as opposed to the average) of $E$ are normally distributed around this value with variance $2\sigma^2$. This implies, that as the size of the system grows, most such matrices will have positive largest eigenvalue, thus making the equilibrium point unstable.

Consider next the case of a non-symmetric matrix, with nondiagonal terms with positive mean, $\mu$, standard deviation, $\sigma$, and whose diagonal terms have mean, $-D$. In this case the largest eigenvalue grows with the size of the matrix as (Furedi and Lomos 1981, Juhasz 1982):

$$E \sim \mu(n-1) - D \tag{9.5}$$

Since $\mu$ is positive these results imply that even if markets are stable when small, they will become unstable as their size becomes large enough for $E$ to change sign.

One argument that could be given for the stability of coupled markets in spite of their size is that not all commodities happen to be coupled to each other. In terms of our theory, this amounts to a *near decomposability* of the Jacobian matrix whereby the entries are such that the further they are from the diagonal the smaller they become (Simon and Ando 1961). Such systems, sometimes called loosely coupled, are very relevant to situations when markets are initially weakly coupled to each other. But as we will now show, even in this case, as the size of the markets increases the equilibrium will be rendered unstable. In terms of the interactions, this situation can describe either the fact that a given item's price is strongly influenced by a few others and weakly by the rest, or a more structured clustering, where the commodities appear in groups (e.g. technology stocks, foreign currencies) wherein their members strongly interact but members of different groups have weak interactions with each other.

Another way of considering this scenario is to imagine initially isolated markets that eventually get coupled through the interaction of mediating interactions, such as roads and communications. In this case the coupling between the initially isolated markets grows in time.

The corresponding matrices for the first case are constructed by selecting off-diagonal entries at random with large magnitude 1 with probability $p$ and small value $\epsilon < 1$ otherwise. In this case $\mu = p + (1-p)\epsilon$ and $\sigma^2 = p(1-p)(1-\epsilon)^2$. As shown in (9.5), the largest eigenvalue will become positive when the system becomes large enough.

In the second case, the commodities are grouped into a hierarchical structure which we assume to be of depth $d$ and average branching ratio, $b$. In this representation, the strength of the interaction between two commodities will decrease based on the number of levels in the hierarchy that one has to climb to reach a large enough common group to which they both belong. Specifically, the interaction strength will be taken to be $R^h$, with $h$ the number of hierarchy levels that separate the two commodities, and $R$ characterizes the reduction in interaction strength that two commodities undergo when they are separated by one further level. The average size of the matrix is given by $n = b^d$ and the mean of the non-diagonal terms can be computed to be

$$\mu = \frac{\sum_{h=1}^{d} b^{h-1}(b-1)R^h}{\sum_{h=1}^{d} b^{h-1}(b-1)} = \frac{(Rb)^d - 1}{b^d - 1}\frac{R(b-1)}{Rb-1} \tag{9.6}$$

In order to represent a reasonable clustering, we need to specify whether the total interaction is dominated by either those neighboring commodities in the hierarchy we choose, or distant ones. In the first case, this amounts to restricting $R$ to be $R < 1/b$. Notice that this choice makes the decreasing interaction strength between commodities overwhelm the increase in their number as higher levels in the hierarchy are considered.

In this situation when $n$ is large Eq. 9.6 becomes

$$\mu = \frac{R(b-1)}{(1-Rb)n}$$

which implies that as $n$ grows the mean goes to zero, leading to a stable system because of (9.5). However, if fluctuations in the coupling strength were taken into account some of the matrices could still become unstable, as was shown above.

In the second case, $R > 1/b$. This implies that in the large $n$ limit the average becomes

$$\mu = R^d \frac{R(b-1)}{Rb-1} = \frac{R(b-1)}{Rb-1} n^{\frac{\ln R}{\ln b}} \tag{9.7}$$

Note that since $-1 < \frac{\ln R}{\ln b} < 0$, $\mu$ goes to zero as the size of the matrix grows, but in slower fashion than the case above. Nevertheless, this subtle difference in convergence to zero amounts to a qualitative difference in the stability of the system. To see this, notice that (9.5) implies that the largest eigenvalue of a random matrix grows as $\mu n$, which increases with $n$ for this value of $\mu$, thus leading to instability when the system gets large enough. The growth in largest eigenvalue with the size of the system is exhibited in Figure 9.1 for a particular choice of

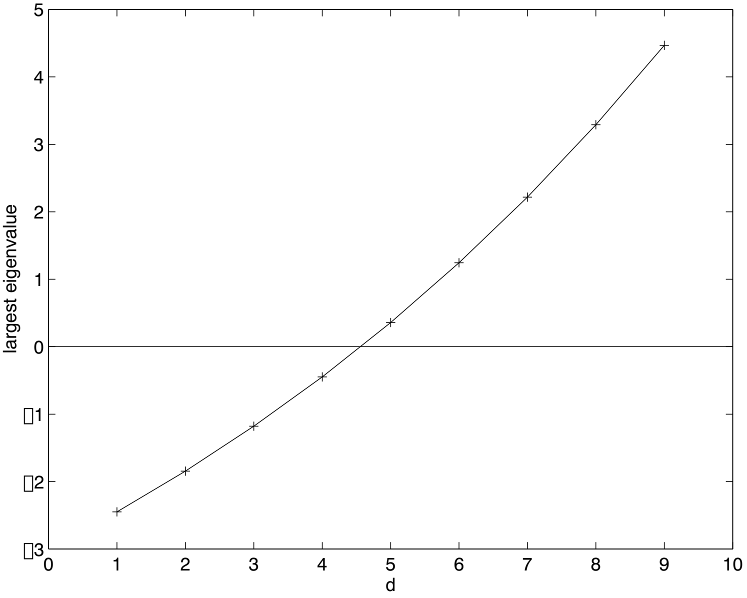


*Figure 9.1* Plot showing the growth of the largest eigenvalue of a hierarchical matrix with branching ratio $b = 2$ and $R = 0.55$ as a function of $d = \log_2 n$. The diagonal elements were chosen to be $D = 3$.The points are the theoretically predicted values of Eq. 9.6 and lie very close to the computed eigenvalues shown by the solid line.

parameters. Notice that the system becomes unstable for $d \geq 5$. Given these results we see that the size of the matrix for which this instability takes place is much larger than the one in the absence of a hierarchy of interactions.

A final possibility is for the commodities to include aggregate structures, such as stock indices. In this case the couplings will correspond to situations where each commodity interacts with itself, the other components of its aggregate, or its higher order aggregate. The ensuing Jacobian will have blocks of nonzero elements and zero entries elsewhere. For this case, we have shown that an even slower growth of the largest eigenvalue with size is obtained (Hogg, Huberman and McGlade 1989). Specifically, the eigenvalues grow no faster than $\sqrt{\ln n}$, implying that much larger coupled markets can be stable when they are structured in a highly clustered fashion. Fluctuation corrections make the eigenvalues grow like $\ln n$, still implying a higher degree of stability than in the previous cases.[1]

## 9.4 Discussion

Recent dramatic fluctuations and losses in the world financial markets have raised concerns about the inherent stability of the global financial system. These concerns have been prompted by the heightened fluidity of global currency flows and the emergence of complicated derivatives, which now allow market players to make financial bets as never before. For example, the derivatives debacle in Orange County, California points to the fact that some market players do not understand the full risks that are being taken. At the same time, the result of the current trends in global finance is that markets are now more and more coupled and individual governments have less and less power to control these perhaps destabilizing couplings. The recent economic crisis in Mexico, the 'tequila effect', resulted in added volatility and 'corrections' in many emerging equity markets all over the world.

In this chapter, we modeled this system as a web of interacting markets much like a biological ecosystem (Rothschild 1990). By doing so we obtained a general result showing that as couplings between previously stable markets grow, the likelihood of instabilities is increased.

These results allow us to understand phenomena that are likely to arise as a system grows in diversity, strength of couplings and in the size of the overall number of the coupled markets. In a sense these results appear to be counterintuitive, for one expects that as a system gets larger, disturbances in a particular part would exhibit a kind of decoherence as they propagate through the system, making it very unlikely that after a given time they would once again concentrate on a particular node and amplify it. But the properties of random matrices make it probable for this conspiracy of perturbations to concentrate on given parts of the market. As the size and diversity of couplings in such matrices grow, the complicated effects of the couplings are more and more likely to result in instability that leads to motion away from equilibrium.

One may ask about the fate of the lost equilibrium and the ensuing evolution of an unstable market. We speculate that once the instability sets in adaptive agents will modify the respective couplings in such a way as to stabilize the entire system once more at another equilibrium. If this is indeed the case, the volatility brought about by instabilities in such large systems is the natural mode by which couplings are modified to achieve a more efficient market. Due to the lack of information needed for appropriate centralized control, it is also by no means clear that regulatory approaches to controlling the market structures will be effective. Improper controls could introduce additional couplings in such an uncertain way that the system may be further destabilized. Accepting the instabilities of these larger systems as the best way for market participants to learn the correct couplings may, therefore, be the most reasonable course of action.

Last but not least, these results cast light on the related problem of the dynamics of multiagent systems in distributed computer networks, which have been shown to behave like economic systems (Huberman and Glance 1995). In the case of only two resources their equilibrium is punctuated by bursts of clustered volatility (Youssefmir and Huberman 1997), a fact which renders the notion of equilibrium suspect. This paper shows that as distributed computing systems get large and more coupled, they will also exhibit a loss of stability, on their way to finding a new and more efficient equilibrium.

## Notes

1 A further interesting possibility (Trefethen *et al* 1993) is that the linearized system can produce an initially growing transient even when $E < 0$ so the perturbations eventually decay. During this transient growth, the perturbations may become large enough to be sustained by nonlinear corrections, thus giving another possible source of instability.

## References

Arrow, K. J. and F. H. Hahn (1971) *General Competitive Analysis*. Holden-Day, San Francisco.

Arrow, K. J. (1988) 'Workshop on the economy as an evolving complex system: Summary', in P. W. Anderson, K. J. Arrow, and D. Pines, (eds) *The Economy as an Evolving Complex System*, 275–81. Addison-Wesley.

Arrow K. J. and Leonid Hurwicz (1958) 'On the stability of competitive equilibrium.' *Econometrica*, 26:522–52.

Brock, W. A. (1988) 'Nonlinearity and complex dynamics and economics and finance.' in P. W. Anderson, K. J. Arrow, and D. Pines, (eds), *The Economy as an Evolving Complex System*, 77–97. Addison-Wesley.

Cohen, J. E. and C. M. Newman (1984) 'The stability of large random matrices and their products.' *Annals of Probability*, 12, 283–310.

Day, R. H. (1975) 'Adaptive processes and economic theory.' in R. H. Day and T. Groves, (eds), *Symposium on Adaptive Economics*, 1–35. Academic Press.

Edelman, A. (1989) *Eigenvalues and Condition Numbers of Random Matrices*. PhD thesis, MIT, Cambridge, MA 02139, May.

Furedi, Z. and K. Komlos (1981) 'The eigenvalues of random symmetric matrices.' *Combinatorica*, 1, 233–41.

Hogg, T. B. A. Huberman and Jacqueline M. McGlade (1989) 'The stability of ecosystems' *Proc. of the Royal Society of London*, B237, 43–51.

Huberman, B. A. and Natalie S. Glance (1995) 'The dynamics of collective action.' *Computational Economics*, 8,27–46.

Juhasz, F. (1982) 'On the asymptotic behavior of the spectra of non-symmetric random (0,1) matrices' *Discrete Mathematics*, 41, 161–5.

May, R. M. (1972) 'Will a large complex system be stable?' *Nature*, 238, 413–14.

McMurtrie, R. E. (1975) 'Determinants of stability of large randomly connected systems' *J. Theor. Biol.*, 50, 1–11.

Mehta,. M. L. (1967) *Random Matrices and the Statistical Theory of Energy Levels*, Academic Press, New York.

Metzler, L. A. (1945) 'Stability of multiple markets: The Hicks condition' *Econometrica*, 13, 277–92.

Nelson, R. A. and S. G. Winter (1982) *An Evolutionary Theory of Economic Change*, Harvard University Press.

Quirk, J. (1981) 'Qualitative stability of matrices and economic theory: A survey article.' in H. J. Greenberg and J. S. Maybee, (eds), *Computer-Assisted Analysis and Model Simulation*, 113–64. Academic Press, NY.

Rothschild, M. L. (1990) *Bionomics: Economy as Ecosystem*. Henry Holt and Company.

Samuelson, P. A. (1941) 'The stability of equilibrium: Comparative statics and dynamics.' *Econometrica*, 9, 111.

Simon, H. A. and Albert Ando (1961) 'Aggregation of variables in dynamic systems.' *Econometrica*, 29, 111–38.

Trefethen, L. N., A. E. Trefethen, S. C. Reddy, and T. A. Driscoll (1993) 'Hydrodynamic stability without eigenvalues.' *Science*, 261, 578–84, July 30.

Wigner, E. P. (1958) 'On the distribution of the roots of certain symmetric matrices.' *Annals Math*, 67, 325–7.

Youssefmir, M.and B. A. Huberman (1997) 'Clustered volatility in multiagent dynamics.' *J. of Economic Behavior and Organizations*, 914, 101–18.

# 10 Heterogeneity, aggregation and capital market imperfection

*Domenico Delli Gatti and*
*Mauro Gallegati*

> 'there can be many representative firms'
>
> Alfred Marshall, *Correspondence*: III, 377

## 10.1 Introduction

The representative agent framework has a long tradition (Marshall, 1920) and has been one of the most successful tools in economics (Hartley, 1997). [1] It is still the cornerstone of microfoundations in macroeconomics and of aggregation in the classical literature (see e.g. Lucas, 1975; Kydland and Prescott, 1982; Long and Plosser, 1983) because the aggregation process allows any difference between the behaviour of individually optimizing agents and that of aggregate variables to be ruled out.

Despite its success, economists are growing more and more dissatisfied with the representative agent framework[2] for a number of different reasons.

First of all, the set of assumptions necessary to reach exact aggregation in a logically consistent way is impressive.[3] Martel (1996: 128) lists the following:

- homothetic preferences (i.e. linear Engel's curve);
- linearly homogeneous production functions, identical for all firms;
- homogeneous and infinitely divisible commodities and factors of production;
- a common set of prices with constant relative ratios;
- constant distribution of income and endowments over time.

Of course, none of the requirements above matches real world features. Empirical analysis at the disaggregated level has shown that heterogeneity matters and aggregation of heterogeneous agents is of central relevance, since there is systematic evidence of individual differences in economic behaviour (Stoker, 1993). In econometrics, several contributions show that aggregate dynamics and individual heterogeneity are intertwined. On the one hand, neglecting heterogeneity in aggregate equations generates spurious evidence of dynamic structure. On the other hand, aggregation of very simple individual behaviour may lead to aggregate complex dynamics.[4]

Apart from the obvious objections to the Representative Agent Hypothesis (RAH) emerging from the empirical evidence, there are theoretical reasons to reject the use and misuse of the representative agent.

First of all, Arrow, 1951 (and later Sonnenschein, 1972, and Mantel,1976) have shown that the proposition according to which the properties of an aggregate function will reflect those of the individual function has no theoretical foundations.

Second, ignoring heterogeneity may lead to a fallacy of composition (see Weiss, 1967) which may be relevant when agents' decisions are not perfectly synchronized – so that composition effects arise (Caballero, 1992) – and misleading for empirical analysis.[5]

Third, the representative agent framework can be unsatisfactory for a wide range of economists, even belonging to opposite sides of the academic spectrum. The very idea of asymmetric information of the new keynesian economics (Mankiw and Romer, 1991) is inconsistent with the representative agent hypothesis (Stiglitz, 1992). In fact, the asymmetric information context is based on heterogeneity; only if agents are different from one another may phenomena such as adverse selection or moral hazard exist.

From a different theoretical perspective, the general equilibrium theorist may not feel at ease with the representative agent assumption because some of the building blocks of general equilibrium theory do not hold in the presence of the representative agent (e.g. the 'Weak Axiom of Revealed Preferences' or Arrow's 'Impossibility Theorem', Kirman, 1992: 122).

Therefore there are theoretical as well as empirical reasons to question the reliability of the representative agent as an economic tool. This chapter reviews some macroeconomic models in which agents' heterogeneity is the main ingredient.

The literature on heterogeneous agents is burgeoning. We can distinguish a number of different strands of this literature: models of distribution and growth,[6] employment and aggregate supply, [7] aggregate demand, [8] and macrodynamics,[9] capital market imperfections and business fluctuations.[10] Some authors identify a law of motion of the average level of the state variable, keeping constant the distribution; others analyze the dynamics of distribution alone, making inferences on aggregate dynamics.

The assumption of a constant distribution of income, wealth or endowments over time is quite implausible when coping with dynamics. In a sense, agents' heterogeneity is the logical requirement for dynamic analysis. According to the impulse-propagation approach small shocks may lead to large fluctuations once a sufficiently large amplification mechanism is activated. If agents are heterogeneous, the propagation mechanism is affected by changes in the distribution of agents. In such a case, business cycles are *not* 'all alike' (Lucas, 1977). Moreover, if we give up the RAH, fluctuations and cycles may emerge as the result of changes in the distribution of agents rather than the consequence of aggregate (as well as idiosyncratic) exogenous shocks to which a 'representative' individual reacts, as in the standard literature on the business cycle. In other words, if the RAH is abandoned, the change in distribution may produce endogenous fluctuations. The literature we review in the next sections emphasizes these aspects.

In the following we survey extensively a class of models in which heterogeneity and aggregation play a crucial role. We deal with macroeconomic models based on capital market imperfections due to asymmetric information. Greenwald and Stiglitz, 1988, 1990, 1993 (section 2), Bernanke and Gertler, 1989, 1990 (section 3), Kiyotaki and Moore, 1997 (section 4), have put forward an approach which emphasizes the role of financial fragility through the interaction of heterogenous agents. These authors, however, don't push the analysis to the point of identifying the law governing the distribution of agents. On the other hand, models by Galor and Zeira, 1993, and Aghion and Bolton, 1997 (section 5) emphasize how the distribution of income and wealth across agents affects growth, i.e. the dynamic behaviour of the system. Section 10.6 discusses an approach which is at the crossroad of the existing literature: while providing an aggregation mechanism, it identifies the law of motion of agents' distribution and shows how aggregation of heterogenous individuals affects the dynamics. Section 10.7 concludes.

## 10.2 Bankruptcy costs, net worth and employment

In a series of papers, Greenwald and Stiglitz (GS hereafter) have put forward a theory of supply decisions in which the financial structure of the firm plays a relevant role (Greenwald and Stiglitz, 1988, 1990, 1993). Their theoretical framework is based on the assumption that firms are unable to raise external finance by issuing new equities because of asymmetric information between managers and potential shareholders on the value of the firm *(equity rationing).* Due to this financing constraint firms rely first and foremost on internal funds in order to finance production and resort to bank credit in order to raise external finance if internal funds are insufficient.

As a consequence, the firm runs the risk of bankruptcy, which occurs when the realization of the random individual price happens to be lower than average operating cost, which in this framework coincides with debt commitments. The probability of bankruptcy is a decreasing function of firms' *net worth* (or *equity base*).

In this framework, the scale of production of each firm depends upon its net worth. In fact, the higher is net worth, the lower is the probability of bankruptcy and the higher employment and production. As a consequence, the risk-neutral firm which takes explicitly into account the cost of bankruptcy behaves *as if* it were risk averse. Production in the presence of bankruptcy risk, in fact, is smaller than in the case in which the firm does not face financing constraints, i.e. capital markets are perfect and the Modigliani–Miller theorem holds true.

It is worth noting that the representative agent assumption is perfectly consistent with the Modigliani–Miller perfect capital markets world which delivers the first best solution. In this case, in fact, individual financial conditions – albeit different from one agent to another – are irrelevant for employment and production decisions. When agents are different, however, the dispersion of financial conditions may generate composition effects and affects the dynamics.

### *10.2.1 Background assumptions*

GS consider an economy populated by firms, households and banks. The production function of the $i$-th firm in period t is $Y_t^i = F(N_t^i)$ where $Y_t^i$ is output, $N_t^i$ employment, and $F(.)$ is a well behaved production function [11]uniform across firms. Therefore $N_t^i = \Phi(Y_t^i)$ is the labour requirement function: $\Phi = F^{-1}$.

GS assume that 'production takes time'. Therefore there will be a one-period time lag between production and sale of the good. Because of the production lag, the price $P_{t+1}^i$ at which the firm will sell in t+1 output produced in t is uncertain. Uncertainty is captured by assuming that $P_{t+1}^i$ is a random variable with expected value equal to the general (average) price level $P_{t+1}$. As a consequence the *relative price* $u_{t+1}^i = P_{t+1}^i / P_{t+1}$ is also a random variable, with c.d.f. $G(.)$ and expected value $E(u_{t+1}^i) = 1$.

Firms finance production costs, i.e. the wage bill, at least partially by means of internally generated funds, which will be referred to as *net worth*. A financing gap occurs when net worth is not sufficient to pay the wage bill. In this case the firm has to resort to external finance. In principle there are two sources of external finance: bank loans and the issue of new equities. GS assume that the issue of new equities is not feasible because of *equity rationing* (Myers and Majluf, 1984; Greenwald *et al.*, 1984), which is due to asymmetric information between potential shareholders and managers of the equity-issuing firm.

Therefore, the only source of external finance available to the firm is credit.[12] The demand for loans in t is $B_t^i = W_t N_t^i - E_t^i$ where $W_t$ is the nominal wage, and $E_t^i$ is equity base or net worth in nominal terms. The demand for loans in real terms is $b_t^i = w_t N_t^i - A_t^i$ where $b_t^i \equiv \frac{B_t^i}{P_t}, w_t \equiv \frac{W_t}{P_t}$ is the real wage and $a_t^i \equiv \frac{E_t^i}{P_t}$ is the equity base in real terms.

Banks extend credit on demand at the interest rate $i_L$. GS assume that debt is repaid completely in one period (there is no accumulation of debt). In this case, nominal debt commitments for the firm in t+1 are $(1+i_L)B_t^i$.

The firm's profit in nominal terms in t+1 is

$$\Pi_{t+1}^i = P_{t+1}^i Y_t^i - (1+i_L) B_t^i$$

Dividing by $P_{t+1}$ we get profit in real terms:

$$\pi_{t+1}^i = u_{t+1}^i Y_t^i - R b_t^i$$

where $R \equiv (1+r) = (1+i_L)\frac{P_t}{P_{t+1}}$ is the gross real interest rate, which will turn out to be equal to the rate of time preference and exogenous. Recalling that $b_t^i = w_t N_t^i - A_t^i$ and that $N_t^i = \Phi(Y_t^i)$ the expression above can be written as:

$$\pi_{t+1}^i = u_{t+1}^i Y_t^i - R\left[w_t \Phi(Y_t^i) - A_t^i\right]$$

The relative price being stochastic, also profit is a random variable with expected value:

$$E(\pi^i_{t+1}) = Y^i_t - R[w_t\Phi(Y^i_t) - A^i_t]$$

The firm goes bankrupt if $\pi^i_{t+1} < 0$, i.e. if

$$u^i_{t+1} < R[w_t\Phi(Y^i_t) - A^i_t]/Y^i_t \equiv \bar{u}^i_{t+1}$$

In other words the firm goes bankrupt if the relative price in t+1 happens to be lower than a critical threshold $\bar{u}^i_{t+1}$, which coincides with debt service per unit of output. The probability of bankruptcy is[13]

$$\Pr(u^i_{t+1} < \bar{u}^i_{t+1}) = G(\bar{u}^i_{t+1}) = G\left(\frac{R[w_t\Phi(Y^i_t) - A^i_t]}{Y^i_t}\right), \quad G' > 0.$$

In other words, the probability of bankruptcy is increasing with the interest rate and decreasing with the equity base. It is increasing also in the level of output if the elasticity of the input requirement to the level of output is greater than one.

Finally, GS assume there are bankruptcy costs which are increasing with the scale of production:[14] $CB^i_t = c(Y^i_t), c' > 0$.

### *10.2.2 Perfect capital markets: the first best case*

Let's assume, for the sake of discussion, that firms don't incur bankruptcy costs. In this case, the problem of the firm is:

$$\max_{Y^i_t} E(\pi^i_{t+1}) = Y^i_t - R[w_t\Phi(Y^i_t) - A^i_t]$$

Solving the problem yields: $Y^i_t = \Phi'^{-1}\left(\frac{1}{Rw}\right)$.Therefore $N^i_t = F'^{-1}(Rw_t)$. It is easy to see that for each firm output supply (and labour demand) is a decreasing function of the real interest rate and the real wage, which are uniform across firms. In other words employment and production decisions react only to changes of the real interest rate and the real wage. Heterogeneity of financial conditions, albeit present, is irrelevant.

Notice that there are two sources of heterogeneity in this framework. The first one is the random relative price, which can be thought of as an idiosyncratic shock. This source of heterogeneity is irrelevant in this context when GS consider the profit maximization problem of the firm. By assumption, in fact, each firm is risk neutral and its expected relative price is one (expectations are homogeneous and rational). The second source of heterogeneity is the equity base. Firms can differ, in fact, according to their degree of financial robustness as measured by the equity base. In the first best case, however, also this source of heterogeneity is irrelevant because there are no bankruptcy costs.

In this case, individual agents, albeit heterogeneous, may be modeled as identical with no loss of generality. Aggregation would yield an aggregate supply and labour demand functions which keep the basic features of the individual functions.

Let's now turn briefly to the household sector. The representative infinitely lived household maximizes the discounted sum of utilities – which are assumed to be linear in consumption and labour supplied – over an infinite horizon subject to a budget constraint according to which total wealth consists of labour income and the return on assets. According to the solution of this dynamic programming problem:

- in each period, consumption is equal to income (output) of the previous period;[15]
- the real interest rate is equal to the (given and constant) rate of time preference;
- the supply of labour is an increasing function of the real wage only.

Equilibrium on the labour market yields $w_t^{fe} = w^{fe}(R)$ and $N_t^{fe} = N^{fe}(R)$ where the *fe* stands for *full employment*. As a consequence

$$Y_t^{fe} = F\left(N^{fe}(R)\right) = Y^{fe}(R)$$

is full employment output. It is easy to see that full employment output is a decreasing function of the real interest rate, provided certain plausible conditions on the configuration of parameters are met.

The law of motion of the equity base of the $i$-th firm in real terms can be written as follows:

$$A_{t+1}^i = u_{t+1}^i Y_t^i - R\left[w_t \Phi(Y_t^i) - A_t^i\right] - D_{t+1}^i \tag{10.1}$$

where $A_{t+1}^i$, the equity base in period t+1, is equal to retained profit, which in turn is equal to total profit less dividends $D_{t+1}^i$. GS assume that dividends are an increasing function of the equity base $D_{t+1}^i = \mathrm{d}(A_t^i), d' > 0$.

The law of motion of the aggregate equity base can be derived from (10.1) through aggregation. Assuming that the labour market is in equilibrium, the law of motion is:

$$A_{t+1} = Y^{fe}(R) - R\left[w^{fe}(R) N^{fe}(R) - A_t\right] - \mathrm{d}(A_t) \tag{10.2}$$

(10.2) is a non linear difference equation in the state variable $A_t$, the aggregate (average) equity base, parametrized to the real interest rate. The steady state solution(s) will be generically written as follows: $A^* = A^*(R)$. In the steady state the equity base on average depends negatively upon the real interest rate. Under quite general assumptions, the steady state is stable.

### *10.2.3 Imperfect capital markets*

If firms run the risk of bankruptcy and bankruptcy is costly, the problem of the firm must be rewritten as follows:

$$\max_{Y_t^i} E(\pi_{t+1}^i) - CB_t^i \Pr(u_{t+1}^i < \bar{u}_{t+1}^i)$$

After substitution one gets

$$\max_{Y_t^i} Y_t^i - R[w_t\Phi(Y_t^i) - A_t^i] - c(Y_t^i)G\left(\frac{R[w_t\Phi(Y_t^i) - A_t^i]}{Y_t^i}\right)$$

Solving the problem yields:

$$Y_t^i = Y^i(Rw_t, A_t^i).$$

Therefore

$$N_t^i = \Phi\left(Y^i(Rw_t, A_t^i)\right) = N^i(Rw_t, A_t^i)$$

In this case, for each firm output and labour demand are increasing functions of the equity base, given the real interest rate and the real wage, which are uniform across firms.

Notice that firms can differ in their production and employment decisions due to the different degrees of financial robustness as measured by the level of the equity base. This is the only source of heterogeneity which persists once the problem of the firm is solved as it is clear from the formulae above. In the first best case, on the contrary, production and employment decisions were independent from the equity base.

According to GS, aggregation would yield an aggregate supply function and a labour demand function which keep the basic features of the corresponding individual functions. Therefore aggregate (average) output and labour demand are increasing functions of the aggregate (average) equity base, given the real interest rate and the real wage, which are uniform across firms. GS would be perfectly right if the equity base were the same for each and every firm (i.e. if they had assumed the representative agent framework from the beginning) – in which case there would not be any difference between the individual and the average relations – and/or if the individual relations were linear. If financial robustness is different from one firm to the other, however, and if the $Y^i(.)$ function is non-linear, also the higher moments of the distribution of the equity base across agents should be taken into account in the aggregation procedure. In particular, the *variance of the equity base across agents* should show up in the aggregate relation.

If we take GS' aggregation procedure at face value and ignore the higher moments of the distribution, equilibrium on the labour market, yields $w_t^{fe} = w^{fe}(R, A_t), N_t^{fe} = N^{fe}(R, A_t)$. As a consequence $Y_t^{fe} = Y^{fe}(R, A_t)$ is full employment output. If we amend GS' aggregation procedure to take heterogeneity seriously, however, the variance of the equity base would be an important part of the story, it should show up in the full employment formulae for real wage, employment and output and composition effects would be crucial.

The law of motion of the equity base of the $i$-th firm in real terms has been derived above (see eq.(10.1)). The law of motion of the aggregate equity base can be derived from (10.1) through aggregation. Summation and averaging over firms yields :

$$A_{t+1} = Y^{fe}(R, A_t) - R\left[w^{fe}(R, A_t)N^{fe}(R, A_t) - A_t\right] - \mathrm{d}(A_t) \tag{10.3}$$

where, once again, we have borrowed the aggregate formulae from GS, ignoring the variance. (10.3) is a non linear difference equation in the state variable $A_t$, parameterized to the real interest rate, which can yield different types of dynamics. The steady state solution(s) will be generically written as follows: $A^* = A^*(R)$.In the steady state the equity base on average depends upon the real interest rate. As a consequence in the steady state also output and employment depend on the same variables: $N^* = N^{fe}(R, A^*(R)) = N^*(R), \;\; Y^* = Y^{fe}(R, A^*(R)) = Y^*(R)$.

A monetary shock affects the steady state only if unexpected (GS *policy effectiveness proposition*).[16]

If we amend GS' aggregation procedure to take heterogeneity seriously, however, the variance of the equity base would be an important part of the dynamic story too. In fact, (10.3) should revritten as follows:

$$A_{t+1} = Y^{fe}(R, A_t, \sigma_A^2) - R\left[w^{fe}\left(R, A_t, \sigma_A^2\right) N^{fe}\left(R, A_t, \sigma_A^2\right) - A_t\right] - d(A_t)$$

where $\sigma_A^2$ is the variance of the distribution of the equity base in t.

## 10.3 Monitoring costs, entrepreneurs' savings and capital accumulation

Bernanke and Gertler (1989,1990) (BG hereafter) have put forward a different theoretical framework which yields basically the same results as in GS. BG assume that in a principal-agent relationship between borrowers and lenders, characterized by asymmetric information and moral hazard, the latter can assess the return on investment only incurring *monitoring costs*. As a consequence, also in their framework the scale of production depends upon a proxy of borrowers' net worth, namely entrepreneurial savings. In fact, the higher is net worth, the lower is the cost of monitoring firms on the part of banks and the higher the volume of credit extended, investment and production. The scale of production in the presence of costly state verification is smaller than in the perfect capital markets case.

### *10.3.1 Background assumptions*

BG assume that there are overlapping generations of agents who live for two periods. In each generation (and in total population, constant by assumption) there is a constant proportion $\eta$ of entrepreneurs. An entrepreneur is an agent endowed with an investment project. Non-entrepreneurs ($1-\eta$ of each generation), on the contrary, cannot access the investment technology. The distribution of agents according to their nature (entrepreneurs and non-entrepreneurs) is exogenous (unexplained) and constant.

There are two types of goods, output and capital. Output can be consumed, invested – i.e. used up as an input of the investment project – or 'stored'. If stored, each unit of output yields a constant return $r$. In this framework, storage is a synonym for investment in a risk-free (safe) asset whose return is $r$. Alternatively, we can think of stored output as the volume of saving which is lent (loanable funds supplied) at the exogenous interest rate $r$.

Output is produced by means of a constant returns to scale technology which uses capital and labour. The production function in per capita terms is: $y_t = \theta f(k_t)$ where $y_t$ is per capita output, $\theta$ is a technological shock, $f(.)$ is a well behaved production function and $k_t$ is per capita capital. Assuming that capital depreciates completely in one period we can claim that capital 'tomorrow' is equal to the flow of investment carried out 'today'.

The $i$-th entrepreneur in t is characterized by a *degree of inefficiency* $\omega_t^i$, a random variable distributed uniformly with support $[0,1]$. The input requirement for each investment project in t $(x_t^i)$ is an increasing function of the degree of inefficiency: $x_t^i = x(\omega_t^i), x' > 0, x(0) > 0, x(1) = x_{\max}$. The return on investment – i.e. the quantity of capital generated by each investment project – is a discrete random variable $\tilde{k}$, with expected value

$$E\left(\tilde{k}\right) = \pi_b k_b + (1-\pi_b)k_g, \quad k_b < k_g,\ \pi_b > 1/2.$$

$\pi_b$ is the probability of the lower outcome of the investment project, i.e. the probability of the 'bad' state of the world. Alternatively, we can think of $\pi_b$ as the proportion of investment projects undertaken by entrepreneurs which yield the lower outcome.

The return on investment is known to the entrepreneur. Non-entrepreneurs can observe the return on investment only incurring *monitoring costs* equal to $\gamma$ units of capital. Therefore, capital in period t+1 is:

$$k_{t+1} = \left[E\left(\tilde{k}\right) - h\gamma\right] I_t \tag{10.4}$$

where $I$ is the number of investment projects per capita undertaken, while $h$ is the proportion of projects monitored.

Each agent is endowed with 1 unit of labour which is supplied inelastically (there is no disutility of labour). In equilibrium the real wage will be equal to the marginal productivity of labour: $w_t = \theta\left[f(k_t) - k_t f'(k_t)\right]$.

Preferences are such that entrepreneurs save all their labour income when young and consume all the return they receive on their project when old. Therefore the entrepreneur's saving (when young) is $S_t^e = w_t \eta$.

### *10.3.2 Perfect capital markets: the first best case*

If information were perfect, there would not be monitoring costs: $\gamma = 0$. An entrepreneur would carry out his investment project in t if the return on investment, i.e. the product of the price of capital in t+1 times the expected real return of the investment project, $q_{t+1}E\left(\tilde{k}\right)$, were greater than or equal to the opportunity cost of investing, which in turn is equal to the return on storage times the input requirement of the investment project $rx_t^i = rx(\omega_t^i)$:

$$q_{t+1}E\left(\tilde{k}\right) \geq rx(\omega_t^i) \tag{10.5}$$

Therefore, we can detect a critical degree of inefficiency:

$$\bar{\omega}_t = x^{-1}\left(\frac{q_{t+1}E\left(\tilde{k}\right)}{r}\right) \tag{10.6}$$

such that all the entrepreneurs with inefficiency lower (higher) than $\bar{\omega}_t$ will invest (not invest).

Thanks to the assumption according to which $\omega_t^i$ is a random variable distributed uniformly with support $[0,1]$, the critical degree of inefficiency $\bar{\omega}_t$ represents also the proportion of entrepreneurs who invest. Therefore the distribution of agents between entrepreneurs and non-entrepreneurs is exogenous while the distribution of entrepreneurs between investors and non-investors is endogenous. The higher the return on investment (the lower the return on storage), the greater the share of entrepreneurs who invest in the population of entrepreneurs.

At this point of the analysis agents can be classified into three groups:

- entrepreneurs who invest, whose share in total population is $\eta\bar{\omega}_t$,
- entrepreneurs who don't invest $\eta(1-\bar{\omega}_t)$,
- non-entrepreneurs $(1-\eta)$.

The last two groups do not invest and therefore employ their saving in the storage technology. In other words they lend their savings (supply loanable funds) to the entrepreneurs. The distinction between borrowers (entrepreneurs who invest) and lenders (entrepreneurs who don't invest and non-entrepreneurs) is at least partially endogenous.[17] In a sense, in this framework, entrepreneurs who are 'too inefficient' – i.e. entrepreneurs whose degree of inefficiency is $\omega_t^i > \bar{\omega}_t$ – give up investment and join the ranks of non-entrepreneurs in lending their funds to entrepreneurs who are 'enough efficient' to invest – i.e. entrepreneurs whose degree of inefficiency is $\omega^i \leq \bar{\omega}$.

The supply of capital is obtained from (10.4) substituting $I_t = \eta\bar{\omega}_t$ and $\gamma = 0$. We obtain:

$$k_{t+1} = E\left(\tilde{k}\right)\eta\bar{\omega}_t \tag{10.7}$$

Substituting the definition of $\bar{\omega}_t$ into (10.7) and solving for $q_{t+1}$ we get:

$$q_{t+1} = \frac{rx\left(k_{t+1}/E\left(\tilde{k}\right)\eta\right)}{E\left(\tilde{k}\right)} \tag{10.8}$$

(10.8) is the equation of the supply of capital. Assuming, for the sake of simplicity that $E\left(\tilde{k}\right) = 1$, (10.8) specializes to:

$$q_{t+1} = rx\left(k_{t+1}/\eta\right) \tag{10.9}$$

The demand for capital is obtained from the usual condition according to which the marginal productivity of capital must be equal to the real remuneration of capital:

$$q_{t+1} = \theta f'(k_{t+1}) \tag{10.10}$$

Equilibrium in the market for capital yields $k^*_{t+1} = k(\theta)$; $q^*_{t+1} = q(\theta)$. Per capita output is $y^* = \theta f\left(k(\theta)\right)$.

There are no true dynamics in the first best case. Both the price and the quantity of capital (and therefore output produced) depend in each period only on the technological parameter.[18] For instance an exogenous technological innovation would yield a once and for all increase in the price and the quantity of capital.

Moreover, the volume of internal finance (entrepreneurial savings) has no role to play in capital accumulation and production decision. Financial conditions are irrelevant.

### *10.3.3 Imperfect capital markets*

In the presence of asymmetric information between borrower and lender, the former has an incentive to lie to the latter, declaring the bad outcome when the good one has occurred. The lender can ascertain the true return on investment only by incurring monitoring costs. BG derive the optimal financial contract and show that auditing/monitoring occurs only if the bad state of the world is declared. They distinguish between the case of full collateralization and the case of incomplete collateralization.

For the *i*-th entrepreneur, *full collateralization* occurs if:

$$q_{t+1}k_b \geq r\left[x(\omega^i_t) - S^e_t\right] \tag{10.11}$$

i.e. if the return on investment in the bad state of the world is high enough to allow the entrepreneur to reimburse its debt. Manipulating (10.11) one realises that full collateralization occurs if the volume of internal finance is higher than a threshold which is increasing with the degree of inefficiency:

$$S^e_t \geq x(\omega^i_t) - \frac{q_{t+1}k_b}{r} = S(\omega^i_t)$$

In this case the probability of auditing is equal to zero: $p_a = 0$.

*Incomplete collateralization* occurs if:

$$q_{t+1}k_b < r\left[x(\omega^i_t) - S^e_t\right] \tag{10.12}$$

In this case the probability of auditing is positive.

The return on investment in this case is $q_{t+1}\left[E\left(\tilde{k}\right) - \pi_b\gamma p_a\right]$. An entrepreneur with inefficiency $\omega^i_t$ will invest if

$$q_{t+1}\left[E\left(\tilde{k}\right) - \pi_b\gamma p_a\right] \geq rx(\omega^i_t) \tag{10.13}$$

(10.13) replaces (10.5) as a condition for investment to be carried out.

Entrepreneurs with inefficiency higher than $\bar{\omega}_t$ as defined in (10.6) will not invest in the perfect information case. Therefore, they will not invest even in the imperfect information case, whatever their level of internal finance. They will be labelled *poor entrepreneurs* for short. Poor entrepreneurs are lenders.

If the probability of auditing is one, the condition for investment is

$$q_{t+1}\left[E\left(\tilde{k}\right)-\pi_b\gamma\right]\geq rx(\omega_t^i) \tag{10.14}$$

Therefore, we can define a critical degree of inefficiency:

$$\underline{\omega}_t = x^{-1}\left(\frac{q_{t+1}\left[E\left(\tilde{k}\right)-\pi_b\gamma\right]}{r}\right) \tag{10.15}$$

Entrepreneurs with inefficiency lower than $\underline{\omega}_t$ (*good entrepreneurs* for short) will invest even if the probability of auditing is one, whatever their level of internal finance. Fully collateralized good entrepreneurs will never be monitored. As far as incompletely collateralized good entrepreneurs are concerned, BG show that they are monitored with probability

$$p_a = \frac{r\left[x(\omega_t^i)-S_t^e\right]-q_{t+1}k_b}{q_{t+1}\left[(1-\pi_b)(k_g-k_b)-\pi_b\gamma\right]} = p(\omega_t^i, q_{t+1}, S_t^e);\;\; \omega_t^i < \underline{\omega}_t \tag{10.16}$$

The probability of auditing is increasing (decreasing) in the degree of inefficiency (internal finance).

Entrepreneurs with inefficiency higher than $\underline{\omega}_t$ and lower than $\bar{\omega}_t$ will be labelled *fair entrepreneurs* for short. By assumption a fair entrepreneur will invest only if he is fully collateralized, that is if his internal finance is higher than the threshold level of internal finance:

$S_t^e > S(\omega_t^i);\, \underline{\omega}_t < \omega_t^i < \bar{\omega}_t$

or:

$$s(\omega_t^i, q_{t+1}, S_t^e) = \frac{rS_t^e}{rx(\omega_t^i)-q_{t+1}k_b} > 1 \tag{10.17}$$

At this point of the analysis agents can be classified into six groups:

- fully-collateralized good entrepreneurs: they invest, whatever their level of internal finance, and will never be monitored;
- incompletely-collateralized good-entrepreneurs: they invest, whatever their level of internal finance and will be monitored with probability as in (10.16);
- fully-collateralized fair-entrepreneurs: they invest and will never be monitored;
- incompletely-collateralized fair-entrepreneurs: they will not invest;

- poor entrepreneurs;
- non-entrepreneurs.

The taxonomy of agents in the imperfect capital markets case is much richer than in the perfect capital markets case. The first three groups consists of entrepreneurs who invest. Good entrepreneurs, whose share in total entrepreneurial population is $\underline{\omega}_t$, will always invest but the return on investment should be diminished to take into account monitoring costs if the good entrepreneur is not fully collateralized. The probability that a good entrepreneur is not fully collateralized is

$$\int_0^{\underline{\omega}_t} p(\omega_t^i, q_{t+1}, S_t^e)\mathrm{d}\omega_t^i.$$

As to fair entrepreneurs, by assumption only fully collateralized entrepreneurs invest. The share of fully collateralized fair entrepreneurs in total entrepreneurial population is

$$\int_{\underline{\omega}_t}^{\bar{\omega}_t} s(\omega_t^i, q_{t+1}, S_t^e)\mathrm{d}\omega_t^i.$$

The last three groups do not invest and therefore employ their saving supplying loanable funds to the entrepreneurs.

Therefore, the supply of capital will be:

$$\begin{aligned} k_{t+1} &= E\left(\tilde{k}\right)\eta\left[\underline{\omega}_t - \pi_b\gamma\int_0^{\underline{\omega}_t} p(\omega_t^i, q_{t+1}, S_t^e)\mathrm{d}\omega_t^i\right] + \\ &\quad + E\left(\tilde{k}\right)\eta\left[\int_{\underline{\omega}_t}^{\bar{\omega}_t} s(\omega_t^i, q_{t+1}, S_t^e)d\omega_t^i\right] \end{aligned} \tag{10.18}$$

Imposing $E\left(\tilde{k}\right) = 1$ and rearranging we get:

$$k_{t+1} = \bar{\omega}_t - \left[\pi_b\gamma\int_0^{\underline{\omega}_t} p(\omega_t^i, q_{t+1}, S_t^e)d\omega_t^i + \int_{\underline{\omega}_t}^{\bar{\omega}_t}[1 - s(\omega_t^i, q_{t+1}, S_t^e)]\mathrm{d}\omega_t^i\right]\eta \tag{10.19}$$

The supply of capital in the imperfect capital markets case depends not only on the price of capital as in the first best case but also on the volume of entrepreneurial savings.

The demand for capital is obtained from the usual condition according to which the marginal productivity of capital must be equal to the real remuneration of capital (see (10.10) above).

Equilibrium in the market for capital yields[19]:

$$k_{t+1}^{*}=k(\theta,S_{t}^{e});q_{t+1}^{*}=q(\theta,S_{t}^{e})$$

In the imperfect capital markets case, therefore, the price and the quantity of capital are influenced not only by the state of technology as in the first best case but also by the financial conditions of investing entrepreneurs captured by their savings. For instance an increase in entrepreneurial savings would produce an increase in the quantity of capital and a decrease of its price.

We recall that entrepreneurial savings in period t are equal to $S_t^e = w_t\eta$ and that $w_t = \theta\left[f(k_t) - k_t f'(k_t)\right]$. Therefore, the equilibrium stock of capital in period t+1 will be:

$$k_{t+1}=k\left(\theta,\eta\theta\left[f\left(k_{t}\right)-k_{t}f'\left(k_{t}\right)\right]\right) \tag{10.20}$$

(10.20) is a generally non-linear first order difference equation in the stock of capital. In this case, there are true dynamics and the volume of internal finance (entrepreneurial savings, i.e. the wage bill of entrepreneurs who invest) has an important role to play. The steady state of (10.20) will be $k^* = k^*(\theta)$.

## 10.4 Moral hazard, financing constraints and collateralizable wealth

Kiyotaki and Moore (1997) (KM hereafter) assume that in a principal-agent relationship between borrowers and lenders, characterized by asymmetric information and moral hazard, borrowers face a financing constraint: the loans they get are smaller or equal to the *value of their collateralizable assets*, which plays, in this framework a role analogous to that of net worth or the equity base in GS and entrepreneurs' savings (internal finance) in BG. As a consequence, also in their framework production depends upon 'net worth'. In fact, the higher is net worth, the higher is the volume of credit extended, investment and production. Production in the presence of this financing constraint is smaller than in the case in which the firm does not face financing constraints, i.e. the perfect capital markets case.

### *10.4.1 Background assumptions*

KM assume that there are infinitely lived agents. In total population, there is a constant proportion of financially constrained agents ('farmers'). Farmers are agents endowed with inalienable human capital. Therefore, they can get from lenders no more than the value of their collateralizable assets. This is the reason of the financing constraints.[20]

Notice that the financing constraint here is binding. Non-farmers ('gatherers'), on the contrary, are not endowed with inalienable human capital. Therefore they do not face financing constraints. It will turn out that all the farmers (gatherers) will be borrowers (lenders). The distribution of agents according to their nature (financially constrained farmers/borrowers and unconstrained gatherers/lenders) therefore, is exogenous and constant.

There are two types of goods, output ('fruit') and a collateralizable, durable, non-reproducible asset ('land') whose total supply is fixed ($\bar{K}$). Output can be consumed or lent. If lent, each unit of output yields a constant return $R = 1 + r$. Output is produced by means of a technology which uses land and labour.

By assumption farmers and gatherers have access to different technologies.

The production function of each financially constrained agent (farmer) is: $y_t^F = (\alpha + \bar{c})k_{t-1}^F$ where $y_t^F$ is output of the farmer in t, $\alpha, \bar{c}$ are positive technological parameters and $k_{t-1}^F$ is land of the farmer in t-1. $\bar{c}k_{t-1}^F$ is the output which deteriorates ('bruised fruit') and is therefore non-tradable.

Each farmer's technology is idiosyncratic in the sense that once production has started only the farmer has the skills to successfully complete the production process, i.e. to make land bear fruit. If the farmers withdrew their labour, production would not be carried out, i.e. land would bear no fruit. In the words of Hart and Moore (1994), the farmer's human capital is inalienable. As a consequence, if the farmers are indebted, they may have an incentive to threaten their creditors to withdraw their labour and repudiate her debt. Creditors protect themselves against this threat by collateralizing the farmer's land. This is the reason why farmers face a *financing constraint:*

$$b_t^F = \frac{q_{t+1}k_t^F}{R} \tag{10.21}$$

According to (10.21), the maximum amount of debt a farmer succeeds to get 'today' $b_t^F$ is such that the sum of principal and interest $Rb_t^F$ is equal to the value of the farmer's land when the debt is due, i.e. $q_{t+1}k_t^F$ where $q_{t+1}$ is the price of land at time t+1.

Farmers face also a *flow-of-funds constraint*:

$$y_t^F + b_t^F = q_t(k_t^F - k_{t-1}^F) + Rb_{t-1}^F + c_t^F \tag{10.22}$$

where $c_t^F$ is the farmer's consumption. Substituting (10.21) into (10.22) we get the budget constraint:

$$c_t^F = (\alpha + \bar{c})k_{t-1}^F - \mu_t k_t^F \tag{10.23}$$

where $\mu_t = q_t - \frac{q_{t+1}}{R}$.

Preferences are such that farmers consume only non-tradable output, i.e. $c_t^F = \bar{c}k_{t-1}^F$. The farmer's demand for land, therefore, is:

$$k_t^F = \frac{1}{\mu_t}\left[(\alpha + q_t)k_{t-1}^F - Rb_{t-1}^F\right] = \frac{\alpha}{\mu_t}k_{t-1}^F \tag{10.24}$$

Substituting (10.24) into (10.21), we obtain:

$$b_t^F = \frac{q_{t+1}}{R}\frac{1}{\mu_t}\left[(\alpha + q_t)k_{t-1}^F - Rb_{t-1}^F\right] \tag{10.25}$$

The production function of each gatherer is: $y_t^G = f(k_{t-1}^G)$ where $y_t^G$ is output of the gatherer in t, $f(.)$ is a well behaved production function and $k_{t-1}^G$ is land of the gatherer in t-1. The gatherers' human capital is not inalienable. Therefore, gatherers face only a *flow-of-funds constraint*:

$$y_t^G + Rb_{t-1}^G = q_t(k_t^G - k_{t-1}^G) + b_t^G + c_t^G \tag{10.26}$$

Substituting the production function of gatherers and the financing constraint of farmers into (10.26) and assuming, for the sake of simplicity and without loss of generality, that population consists only of one farmer and one gatherer so that $k_t^F = \bar{K} - k_t^G$ we get the budget constraint:

$$c_t^G = f(k_{t-1}^G) + \mu_t \left(\bar{K} - k_t^G\right) \tag{10.27}$$

Preferences of the gatherer are such that $R\mu_t = f'(k_t^G)$. The demand for land, therefore, is:

$$k_t^G = f'^{-1}(R\mu_t) \tag{10.28}$$

### 10.4.2 Perfect capital markets: the first best case

In the absence of asymmetric information, i.e. in the perfect capital markets case, there would not be any financing constraints. In equilibrium, the marginal productivity of land for the farmer should be equal to the marginal productivity of land for the gatherer, i.e.

$$f'(k_t^G) = \alpha + \bar{c} \tag{10.29}$$

From (10.29) it follows that the equilibrium quantity of land for the farmer and the gatherer are:

$$\begin{aligned} k_t^G &= f'^{-1}(\alpha + \bar{c}) \\ k_t^F &= \bar{K} - f'^{-1}(\alpha + \bar{c}) \end{aligned}$$

Moreover, from (10.28) it follows that

$$\mu_t = \frac{f'\left(f'^{-1}(\alpha + \bar{c})\right)}{R}$$

There are no true dynamics in the first best case. The quantity of land allocated to (and therefore output produced by) the farmer and the gatherer respectively depend in each period only on the parameters which characterize the production functions of the two types of agents.

Moreover, the value of collateralizable wealth has no role to play in the process of allocation of land and in production decisions. Once again, in the perfect capital markets case, financial conditions are irrelevant.

### 10.4.3 Imperfect capital markets and dynamics

From (10.28) we know that the following must be true:

$$\bar{K} - k_t^F = f'^{-1}(R\mu_t) \tag{10.30}$$

Substituting this expression into (10.24) and rearranging we end up with the following:

$$k_t^F = \frac{R\alpha}{f'(\bar{K} - k_t^F)} k_{t-1}^F \tag{10.31}$$

(10.31) is a non-linear difference equation in the state variable $k_t^F$.

In the steady state $k_t^F = k_{t-1}^F = k^*$ Therefore $\mu^* = a$. Moreover $q_t = q_{t+1} = q^*$ so that $q^* = \alpha\frac{R}{R-1}$. Finally $b_t^F = b_{t-1}^F = b^*$so that $b^* = \alpha\frac{k^*}{R-1}$ . From the steady state condition $k_t^G = k_{t-1}^G = \bar{K} - k^*$ follows $\bar{K} - k^* = f'^{-1}(R\alpha)$. As a consequence: $k^* = \bar{K} - f'^{-1}(R\alpha)$ .

In this setting KM show that small shocks – for instance to technology – can produce large and persistent fluctuations in output and asset prices. In their model, in fact, the durable, non-reproducible asset (land) plays the dual role of a factor of production for both constrained and unconstrained agents and of collateralizable wealth for financially constrained agents. Therefore the price of assets affects the borrowers' financing constraint. At the same time, the size of the borrowers' credit limits feeds back on asset prices. 'The dynamic interaction between credit limits and asset prices turns out to be a powerful transmission mechanism by which the effects of shocks persist, amplify and spread out' (Kiyotaki and Moore, 1997:212).

## 10.5 Capital market imperfections, distribution and growth

In this section we briefly present and discuss the papers by Galor and Zeira (1993) and Aghion and Bolton (1997)[21] in which the authors describe the evolution over time of the distribution of wealth among individuals in the presence of capital market imperfections. Changes in the distribution of wealth, in turn, affect production decisions and economic growth.

Galor and Zeira (1993) (GZ hereafter) assume there are overlapping generations of agents who live for two periods. In each generation (and in total population, constant by assumption) there is a proportion of agents who are skilled (unskilled). An unskilled agent is endowed with a constant returns to scale technology to produce output $(Y^U)$ by means of labour $(N^U)$in both periods of life. The production function of the unskilled is $Y^U = w_u N^U$ where $w_u$ is the real wage of the unskilled worker. A skilled agent invests in human capital (and does not work) when young to acquire a higher degree of efficiency at work when old. The amount of investment in human capital is fixed $(H)$. The production function of the skilled is $Y^S = f(N^S, K)$ where $K$ is the capital stock, $f(.)$ is a well behaved production function, $w_s$ is the real wage of the skilled worker, $w_s > w_u$ . The distribution of agents according to their nature (skilled and unskilled) is endogenous, as we will see.

Each agent has one parent, whom she receives a bequest from, and one child, whom she leaves a bequest to. Preferences are defined over consumption in the second period of life (for the sake of simplicity, young agents do not consume) and bequest to the child. Each agent's wealth is the sum of labour income in the first period of life and initial wealth, i.e. bequest received from the parent.

Capital markets are imperfect. Borrowers can evade debt payments by ‘moving to other places’ and lenders can avoid this occurrence by ‘keeping track’ of borrowers. This is a moral hazard problem in a principal-agent relationship that the principal can deal with by incurring monitoring costs. In other words the basic setting is reminiscent of BG’s. The actual interest rate on loans $(\rho)$ is a mark up on the exogenous ‘basic’ interest rate $(r)$ which GZ think of as the ‘world rate of interest’.

There are three types of individuals in this economy:

- unskilled workers who are lenders;
- skilled workers who inherit less than the value of the investment in human capital. They are borrowers;
- skilled workers who inherit more than the investment in human capital. They are lenders.

GZ assume that preferences are homogeneous across agents. Whatever her type, each agent has a linear logarithmic utility function defined over consumption (when old) and bequest.

From utility maximization follows that the (optimal) bequest each agent leaves to his child at time t – which will be the initial wealth of the newborn in t+1 (say $Z_{t+1}$) – is a fraction $(1-\zeta)$ of her wealth, which in turn is the sum of labour income in the first period of life and initial wealth.

The definition of agents’s wealth however is not uniform across agents. Therefore the bequest received from her parent by each child can be described by the following laws of motion:

$$Z_{t+1} = (1-\zeta)\left[(1+r)\left(Z_t + w_u\right) + w_u\right] \tag{10.32}$$

$$Z_{t+1} = (1-\zeta)\left[-(1+\rho)\left(H - Z_t\right) + w_s\right]; \;\; Z_t < H \tag{10.33}$$

$$Z_{t+1} = (1-\zeta)\left[(1+r)\left(Z_t - H\right) + w_s\right]; \;\; Z_t > H \tag{10.34}$$

$(10.32)$ is the bequest received by the child of an unskilled worker. The unskilled worker’s wealth is the sum of labour income when old and the gross return on her saving when young, which in turn is the sum of initial wealth and labour income when young augmented by the flow of interest on loans. (10.33) is the bequest received by the child of a skilled worker who has to borrow in order to carry out her investment in human capital. (10.34) is the bequest received by the child of a skilled worker who lends the difference between her initial wealth and her investment in human capital.

The system (10.32), (10.33), (10.34) generates a piecewise linear phase diagram with two steady states, $Z_u^*$ and $Z_s^*$, $Z_u^* < Z_s^*$, which are both stable. Let’s assume that there is an exogenously given initial distribution of wealth across individuals. For the sake of simplicity, we can assume that individual wealth is distributed uniformly with support $[0, Z_{\max}]$.If an individual has initial wealth smaller (greater) than a critical degree of wealth $\bar{Z}$, she leaves a bequest to her child smaller (greater) than her initial wealth and her child would do the same, so that with the passing of time the individual wealth of her descendants will converge to $Z_u^*$ ( $Z_s^*$). The distribution of wealth, in other words, will converge to a *stationary distribution* in which a portion of total population$(\bar{Z}/Z_{\max})$ has wealth equal to $Z_u^*$ while the remaining $\left(1 - \frac{\bar{Z}}{Z_{\max}}\right)$has wealth equal to $Z_s^*$. Per capita wealth is therefore:

$$z = Z_s^* - \frac{\bar{Z}}{Z_{\max}} \left(Z_s^* - Z_u^*\right)$$

This is a simple theory of the persistence of heterogeneity. If the dynamics of the distribution were such as to converge to a stationary distribution in which only one type of agents survive, we would have had the restoration of the representative agent. This is what would happen in case of perfect capital markets.

Aghion and Bolton (1997) (AB hereafter) assume that agents live for one period during which they work and invest. Income earned by working and investing is divided between consumption and bequests. Each agent has one parent, whom she receives a bequest from, and one child, whom she leaves a bequest to. Each agent's wealth is equal to the bequest received from the parent. The distribution of wealth endowments will be represented by $G(Z)$.

Each agent is endowed with one unit of labour which she supplies inelastically (there is no disutility of labour). She can use her unit of labour to work on a 'backyard activity' which yields $\bar{r}$ with certainty or invest in an 'entrepreneurial activity' which yields $r$ with probability $p_r$ or 0 with probability $(1 - p_r)$.

In order to invest, the agent must commit one unit of wealth. If the agent doesn't invest, she can employ her wealth in an economy-wide 'mutual fund'.

Capital markets are imperfect. The borrower–lender relationship is characterized by a moral hazard problem with limited wealth constraints. In this setting AB show that, due to the features of the optimal lending contract, the probability of success $p_r$ and the interest rate charged on loans $\rho$ are functions of individual wealth, with $\frac{\partial p_r}{\partial Z} > 0, \frac{\partial \rho}{\partial Z} < 0$.

In equilibrium the expected return on loans must be uniform across borrowers and equal to the interest rate of the mutual fund $r_m$, i.e. $r_m = p_r(Z)\rho(Z)$

An agent endowed with less than one unit of wealth will borrow the amount $(1 - Z)$ and invest (alternatively: carry on her backyard activity and lend) if the expected net return on investment is greater (smaller) than the opportunity cost of investing, which in turn is equal to the sum of the return on the mutual fund and the return on the backyard activity. This condition entails a critical level of individual wealth $\tilde{Z}$, which is an increasing function of the return on the mutual fund, such that those individuals with initial wealth $Z > \tilde{Z}(r_m)$ (alternatively: $Z < \tilde{Z}(r_m)$) prefer to borrow and invest (work in the backyard and lend). The demand for loans, therefore, will be $B^d := \int_{\tilde{Z}(r_m)}^{1} (1 - Z) dG(Z) = B^d(r_m)$. The supply of loans, on the other hand, is the sum of resources made available by poor individuals who prefer to lend and by rich individuals who hold wealth in excess of the input requirement of investment:

$$B^s := \int_0^{\tilde{Z}(r_m)} Z\,\mathrm{d}G(Z) + \int_1^{\infty} (Z-1)\,\mathrm{d}G(Z) = B^s(r_m)$$

Equilibrium on the credit market, whenever it exists, yields the equilibrium level of the interest rate, say $r_m^*$,which depends only on the features of the distribution function, and the equilibrium distribution of agents into three groups:[22]

- fully-collateralized (rich) entrepreneurs who invest and lend the excess of their wealth over the input requirement of investment;
- incompletely-collateralized (poor) entrepreneurs who invest and borrow the excess of the input requirement of investment over their wealth;
- poor non-entrepreneurs who work in the backyard and lend.

The nature of borrowers and lenders, in other words, is endogenously determined as in BG and GZ.

From utility maximization follows that the (optimal) bequest each agent leaves to her child at time t – which will be the initial wealth of the newborn in t+1 (say $Z_{t+1}$) – is a fraction $(1-\zeta)$ of her wealth, which in turn depends on the type of activity she has carried out and her initial wealth.

The bequest received by each child from her parent when the latter is a poor non-entrepreneur ($Z < \tilde{Z}(r_m)$)can be described by the following law of motion:

$$Z_{t+1} = (1-\zeta)\left(r_m^* Z_t + \bar{r}\right)$$

When the parent is a rich entrepreneur, the law of motion is:

$$Z_{t+1} = (1-\zeta)\left[r + r_m^*(Z_t - 1)\right] \tag{10.35}$$

if the investment project yields the high return $r$, with probability $p_r = p_{\max}$ where $p_{\max}$ is the maximum probability of success.

On the other hand, if the investment project yields the low return 0, with probability $1 - p_{\max}$, the law of motion is:

$$Z_{t+1} = (1-\zeta)\left[r_m^*(Z_t - 1)\right] \tag{10.36}$$

Finally, when the parent is a poor entrepreneur, the law of motion is:

$$Z_{t+1} = (1-\zeta)\left[r - \rho(Z_t)(1 - Z_t)\right] \tag{10.37}$$

if the investment project yields the high return $r$, with probability $p_r = p_r(Z_t)$.

On the other hand, if the investment project yields the low return 0, with probability $1 - p_r(Z_t)$, the law of motion is:

$$Z_{t+1} = 0 \tag{10.38}$$

According to AB the economy will grow until all investment opportunities are exploited. In this case the equilibrium interest rate on the mutual fund is $r_m^* = 1$. Applying the results on convergence for monotonic Markov processes AB show that the distribution of wealth converges to a *stationary (long run) distribution*.

As in GZ, AB have put forward a simple theory of the persistence of heterogeneity in wealth endowments in an imperfect capital markets framework.

## 10.6 The evolution and the persistence of heterogeneity: a proposal

In the search for a satisfactory way to model the evolution and persistence of heterogeneity over time and its role in the transmission and amplification of shocks, we have put forward a simple model (Delli Gatti, Gallegati and Palestrini, 2000) with imperfect capital markets along the lines of Greenwald and Stiglitz.

Our economy is characterized by a large number (say $z$) of firms. Each firm produces a homogenous good by means of a constant returns to scale technology in which capital is the only input:

$$Y_t^i = \nu K_t^i \tag{10.39}$$

where $Y_t^i$ and $K_t^i$ are output and capital of the $i$-th firm in period t, $\nu$ is the output-capital ratio, uniform across firms.

Firms sell their output at an uncertain price because of their limited knowledge of market conditions. In order to capture uncertainty, we model the individual selling price $P_t^i$ as: $P_t^i = u_t^i P_t$ where $u_t^i$ is a random idiosyncratic shock and $P_t$ is the average market price, uniform across firms. We assume that $E\left(u_t^i\right) = 1$. Therefore $E\left(P_t^i\right) = P_t$.We can also interpret the random shock as the relative price: $u_t^i = P_t^i/P_t$.

Moreover we assume that firms cannot raise funds on the Stock market because of equity rationing but they have unlimited access to credit. This means that firms do not issue new equities but can obtain from banks all the credit they need to finance production at the (exogenous) rate of interest, $r$, uniform across firms and time invariant.

Firms differ according to their financial conditions. The financial robustness of a firm is proxied by the *equity ratio*, i.e. the ratio of its equity base to capital $a_t^i = A_t^i/K_t^i$.

Each firm incurs financing costs $CF_t^i$ equal to debt commitments:

$$CF_t^i = r(qK_t^i - A_{t-1}^i) \tag{10.40}$$

where $q$ is the real price of capital and $A_{t-1}^i$ is the net worth or equity base in real terms inherited from the past.

Assuming that there is no depreciation, capital accumulates according to the investment equation $I_t^i = K_t^i - K_{t-1}^i$.

We assume that the firm incurs capital adjustment costs, $CA_t^i$ which are increasing in the ratio $I_t^i/K_t$of its investment $I_t^i$to the *average* capital stock$K_t$. We make the additional technical assumption that the adjustment costs function is quadratic:

$$CA_t^i = \frac{\gamma}{2}\frac{\left(I_t^i\right)^2}{K_t} = \frac{\gamma}{2}\frac{\left(K_t^i - K_{t-1}^i\right)^2}{K_t} \tag{10.41}$$

Quadratic adjustment costs are well known in the literature on investment. The novelty of the expression above consists in assuming that adjustment costs are decreasing with the average capital stock: the higher is the average capital stock – i.e. investment activity on the part of other firms – the lower is the level of adjustment costs for the individual firm. This formulation captures an *externality* in investment activity.

Real profit is the difference between real revenue and real cost, which in turn is the sum of financing and adjustment costs:

$$\begin{aligned} \pi_t^i &= u_t^i Y_t^i - CF_t^i - CA_t^i = \\ &= u_t^i Y_t^i - r(qK_t^i - A_{t-1}^i) - \frac{\gamma}{2}\frac{\left(K_t^i - K_{t-1}^i\right)^2}{K_t} \end{aligned}$$

In this framework, bankruptcy occurs if net worth becomes negative. Net worth 'today' is equal to net worth 'yesterday' plus retained profit, which in turn is equal to profit less the flow of dividends ($D_t^i$). The bankruptcy condition therefore is:

$$A_t^i = A_{t-1}^i + \pi_t^i - D_t^i = A_{t-1}^i + u^i Y_t^i - CF_t^i - CA_t^i - D_t^i < 0 \tag{10.42}$$

The inequality (10.42) is verified if the sum of financing costs, adjustment costs and dividends is higher than revenues – thereby generating a loss – and the associated loss is higher than the equity base inherited from the past. In order to simplify the argument, in the following we will assume that the flow of dividends is proportional to net worth inherited from the past:

$$D_t^i = rA_{t-1}^i \tag{10.43}$$

In a sense this is tantamount to assuming the net worth is remunerated at the same rate as bank loans.

Substituting(10.40)(10.41)(10.43) into (10.42) we get:

$$A_{t-1}^i + u_t^i Y_t^i - r(qK_t^i - A_{t-1}^i) - \frac{\gamma}{2}\frac{\left(K_t^i - K_{t-1}^i\right)^2}{K_t} - rA_{t-1}^i < 0$$

Using (10.39) and rearranging, we can write the bankruptcy condition as follows:

$$u_t^i < r\frac{q}{\nu} - \frac{A_{t-1}^i}{\nu K_t^i} + \frac{\gamma}{2}\frac{\left(K_t^i - K_{t-1}^i\right)^2}{\nu K_t^i K_t} \tag{10.44}$$

In the following, for the sake of analytical convenience, we will adopt a simplified version of the bankruptcy condition, namely:

$$u^i < r\frac{q}{\nu} - \frac{A_{t-1}^i}{\nu K_t^i} \equiv \bar{u}_t^i \tag{10.45}$$

The condition (10.45) essentially ignores capital adjustment costs. This assumption makes computations much less cumbersome without loss of generality. According to the bankruptcy condition, bankruptcy occurs if the realization of the random relative price $u_t^i$ falls below a critical threshold $\bar{u}_t^i$ which in turn is a function, among other variables, of the capital stock and of the equity base lagged one period.

Let's assume, for the sake of simplicity, that $u_t^i$ is a uniform random variable, with support (0,2). In this case, the probability of bankruptcy is:

$$\Pr\left(u_t^i < \bar{u}_t^i\right) = \frac{\bar{u}_t^i}{2} = \frac{1}{2}\left(r\frac{q}{\nu} - \frac{A_{t-1}^i}{\nu K_t^i}\right)$$

Following Greenwald and Stiglitz, moreover, we assume that bankruptcy is costly (Gordon and Malkiel, 1981; Altman, 1984; Gilson, 1990; Kaplan and Reishus, 1990). In particular, bankruptcy costs, $CB_t^i$ are a decreasing function of the degree of financial robustness, proxied by the equity ratio of the previous period and an increasing function of output:

$$CB_t^i = \left(\alpha_1 - \alpha_2 a_{t-1}^i\right) Y_t^i \tag{10.46}$$

where $\alpha_1$ and $\alpha_2$ are positive parameters.

In this setting, the problem of the firm is:

$$\max E(\pi_t^i) - CB_t^i \Pr(u_t^i < \bar{u}_t^i)$$

After substitution, we can reformulate the firm's problem as:

$$\max_{K_t^i} \nu K_t^i - r^i K_t^i - \frac{\gamma}{2}\frac{\left(K_t^i - K_{t-1}^i\right)^2}{K_t} + T$$

where

$$r^i = rq\left(1 + \frac{\alpha_1}{2} - \frac{\alpha_2}{2} a_{t-1}^i\right)$$

$$T = A_{t-1}^i\left(r + \frac{\alpha_1}{2} - \frac{\alpha_2}{2} a_{t-1}^i\right)$$

$r^i$ is the bankruptcy cost augmented interest rate. It is determined as a mark-up over the interest rate charged by banks on loans. This mark-up is a decreasing function of the equity ratio: the higher the equity ratio, i.e. the financial robustness of the firm, the lower the bankruptcy cost augmented interest rate. In a sense, this mark-up captures the idea of the risk of the borrower. Solving the problem yields:

$$\frac{I_t^i}{K_t} = \frac{K_t^i - K_{t-1}^i}{K_t} = \frac{\nu - r^i}{\gamma} = T_0 + T_1 a_{t-1}^i$$

where

$$T_0 \equiv \frac{\nu}{\gamma} - r\frac{q}{\gamma}\left(1 + \frac{\alpha_1}{2}\right)$$

$$T_1 \equiv \frac{rq\alpha_2}{\gamma 2}$$

Since the issue of new equities is ruled out by assumption, in this framework each firm can increase its net worth inasmuch as it accumulates internal funds. The change of the equity base, therefore, coincides with retained profits.

The law of motion of the equity base of the *i*-th firm is:

$$A_t^i = A_{t-1}^i + u_t^i Y_t^i - rqK_t^i - \frac{\gamma}{2}\frac{\left(K_t^i - K_{t-1}^i\right)^2}{K_t} \tag{10.47}$$

Dividing by the individual capital stock, we can derive the law of motion of the equity ratio:

$$a_t^i = a_{t-1}^i \frac{K_{t-1}^i}{K_t^i} + u_t^i \nu - rq - \frac{\gamma}{2}\frac{\left(K_t^i - K_{t-1}^i\right)^2}{K_t K_t^i}$$

The expression $K_{t-1}^i / K_t^i$can be written as:

$$\frac{K_{t-1}^i}{K_t^i} = 1 - \frac{I_t^i}{K_t}\frac{K_t}{K_t^i}$$

In order to make the analysis manageable, we assume that the ratio $\frac{K_t^i}{K_t} \cong 1$. Thanks to this simplifying shortcut we get:

$$\frac{K_{t-1}^i}{K_t^i} = 1 - \frac{I_t^i}{K_t} = 1 - \left(T_0 + T_1 a_{t-1}^i\right)$$

Moreover, thanks to the assumption $\frac{K_t^i}{K_t} \cong 1$ the expression

$$\frac{\left(K_t^i - K_{t-1}^i\right)^2}{K_t K_t^i} = \frac{\left(I_t^i\right)^2}{K_t K_t^i}$$

boils down to:

$$\left(\frac{I_t^i}{K_t}\right)^2 = \left(T_0 + T_1 a_{t-1}^i\right)^2$$

As a consequence, after substitution, the law of motion of the individual equity ratio becomes:

$$a_t^i = a_{t-1}^i \left[1 - \left(T_0 + T_1 a_{t-1}^i\right)\right] + u_t^i \nu - rq - \frac{\gamma}{2}\left(T_0 + T_1 a_{t-1}^i\right)^2$$

After some algebraic manipulation, we get:

$$a_t = \Gamma_1 a_{t-1} - \Gamma_2 a_{t-1}^2 + \Gamma_0^i \tag{10.48}$$

where

$$\Gamma_1 = 1 - T_0\left(1 + \gamma T_1\right)$$

$$\Gamma_2 = T_1\left(1+\frac{\gamma}{2}T_1\right)$$

$$\Gamma_0^i = u_t^i\nu - rq - \frac{\gamma}{2}(T_0)^2$$

From (10.48) through aggregation, assuming that the law of large numbers holds true, we obtain the law of motion of the average equity ratio:

$$a_t = \Gamma_1 a_{t-1} - \Gamma_2 a_{t-1}^2 + \Gamma_0 - \Gamma_2 V_{t-1} \qquad (10.49)$$

where $V_{t-1}$ is the variance of the distribution of the equity ratio in t-1. From the definition of the variance at time t, we derive the following:

$$\begin{aligned} V_t = {} & \Gamma_2^2(\beta_{t-1}-1)V_{t-1}^2 + 2\Gamma_1\Gamma_0\mu_{t-1}^3 + 4\Gamma_2^2 a_{t-1}^2 V_{t-1} \\ & -4\Gamma_1\Gamma_2 a_{t-1}V_{t-1} + \Gamma_1^2 V_{t-1} + 4\Gamma_2^2\mu_{t-1}^3 a_{t-1} \end{aligned}$$

where $\beta_{t-1}$ is a parameter capturing the kurtosis of the distribution, while $\mu_{t-1}^3$ is the third moment from the mean.

(10.49) and (10.50) is a system of two non-linear difference equations in the state variables $a_t$ and $V_t$ which describes the evolution over time of the first two moments of the distribution of the equity ratio. This dynamical system is described by a two-dimensional non-linear map which yields multiple equilibria (steady states), with different dynamical properties depending upon the chosen configuration of parameters.[23]

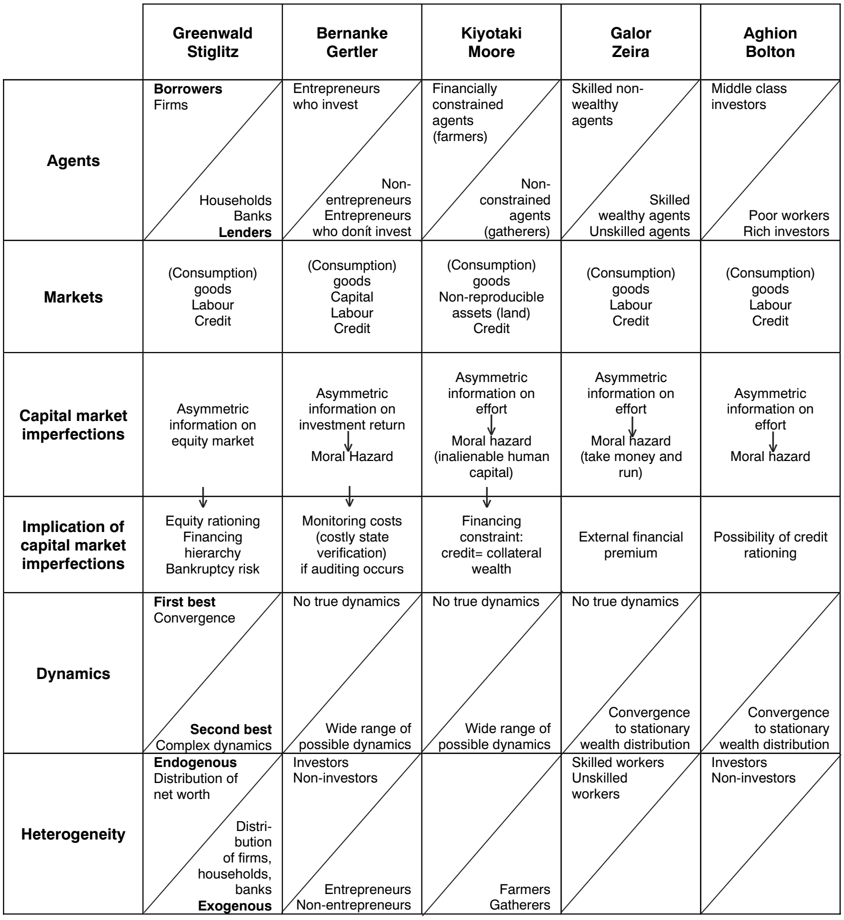
The map generates a wide range of dynamic patterns: convergence to a stationary distribution, periodic or aperiodic cycles, chaotic dynamics, divergence. To each dynamic pattern of the distribution corresponds a dynamic pattern of aggregate production and investment (capital accumulation). Therefore, the framework can generate convergence to a steady growth path, endogenous regular or irregular business cycles and growth – which we label fluctuating growth for short – divergent trajectories, which we label ‘ financial crises’ .[24]

A fluctuation can be generated also by an exogenous stochastic shock forced upon the system when it is in a steady state position. In particular, elsewhere we have shown that the higher the degree of heterogeneity, the larger the effects of a shock and the longer its persistence.

Heterogeneity is very important in business cycle analysis for empirical as well as theoretical reasons (Kirman, 1999). When firms are heterogeneous, knowledge about the distribution of firms is crucial in order to understand the response of the system to aggregate and idiosyncratic shocks.

## 10.7 Conclusion

Heterogeneity is a catch word which can mean different things to different people in different contexts. We have to refine the notion in order to draw meaningful conclusions on its importance in macroeconomics. In this section we will try such a refinement with the help of the following table, which summarizes the main features of the different theoretical frameworks examined so far. The columns are



| | **Greenwald Stiglitz** | **Bernanke Gertler** | **Kiyotaki Moore** | **Galor Zeira** | **Aghion Bolton** |
|---|---|---|---|---|---|
| **Agents** | **Borrowers**<br>Firms<br>Households<br>Banks<br>**Lenders** | Entrepreneurs who invest<br>Non-entrepreneurs<br>Entrepreneurs who donít invest | Financially constrained agents (farmers)<br>Non-constrained agents (gatherers) | Skilled non-wealthy agents<br>Skilled wealthy agents<br>Unskilled agents | Middle class investors<br>Poor workers<br>Rich investors |
| **Markets** | (Consumption) goods<br>Labour<br>Credit | (Consumption) goods<br>Capital<br>Labour<br>Credit | (Consumption) goods<br>Non-reproducible assets (land)<br>Credit | (Consumption) goods<br>Labour<br>Credit | (Consumption) goods<br>Labour<br>Credit |
| **Capital market imperfections** | Asymmetric information on equity market<br>↓ | Asymmetric information on investment return<br>↓<br>Moral Hazard<br>↓ | Asymmetric information on effort<br>↓<br>Moral hazard (inalienable human capital)<br>↓ | Asymmetric information on effort<br>↓<br>Moral hazard (take money and run) | Asymmetric information on effort<br>↓<br>Moral hazard |
| **Implication of capital market imperfections** | Equity rationing<br>Financing hierarchy<br>Bankruptcy risk | Monitoring costs (costly state verification) if auditing occurs | Financing constraint: credit= collateral wealth | External financial premium | Possibility of credit rationing |
| **Dynamics** | **First best**<br>Convergence<br>**Second best**<br>Complex dynamics | No true dynamics<br>Wide range of possible dynamics | No true dynamics<br>Wide range of possible dynamics | No true dynamics<br>Convergence to stationary wealth distribution | Convergence to stationary wealth distribution |
| **Heterogeneity** | **Endogenous**<br>Distribution of net worth<br>Distribution of firms, households, banks<br>**Exogenous** | Investors<br>Non-investors<br>Entrepreneurs<br>Non-entrepreneurs | Farmers<br>Gatherers | Skilled workers<br>Unskilled workers | Investors<br>Non-investors |

*Figure 10.1* Main features of theoretical frameworks examined

entitled to the models presented, while the rows are devoted to a classification of agents (grouped in the two broad categories of borrowers and lenders), markets, type and implications of capital market imperfections, type of dynamics, nature of heterogeneity.

First of all, we can draw a distinction between weak and strong heterogeneity (Rios-Rull, 1995; Gaffeo, 1999). Heterogeneity is *strong* if agents, differentiated according to some economically relevant characteristics, interact strategically. In this case the benchmark theoretical framework is game theory. Heterogeneity is *weak* if agents, albeit different, are not aware of – or simply do not take into account – the fact that the decision taken by each one of them affects the payoff of the others. The models surveyed so far explore the macroeconomic implications of weak forms of heterogeneity. In other words none of the models reviewed in this chapter is embedded in a game theoretical framework.

A second distinction can be drawn between exogenous and endogenous heterogeneity. Heterogeneity is *exogenous* if the distribution of agents according to some economically relevant characteristics is postulated by the modeler, it is *endogenous* if the distribution if determined in the model. An example can help in clarifying this distinction.

A fundamental differentiation of agents in models with imperfect capital markets is the distinction between borrowers and lenders. GS simply assume that distinction from the beginning of the analysis. They start, in fact, from the fact that some agents specialize in lending funds to other agents who specialize in borrowing; they don't bother to assess which characteristics of the agents brought about that specialization. In a sense, the same can be said of KM. In their framework, the distribution between financially constrained and unconstrained agents is exogenous. It turns out that the former are borrowers and the latter are lenders. In this sense also the distribution between borrowers and lenders is exogenous.

On the contrary BG exogenously assume the distribution of agents between entrepreneurs and non-entrepreneurs but are able to derive the distribution of entrepreneurs in investors and non-investors endogenously. It turns out that entrepreneurs characterized by a degree of inefficiency lower than a certain – endogenously determined – threshold level of inefficiency carry out their investment projects and therefore are seeking funds to finance investment activity, i.e. they specialize in investing and borrowing, while entrepreneurs with a degree of inefficiency higher than the above mentioned threshold give up investment and 'store'. In other words, inefficient entrepreneurs and non-entrepreneurs specialize in lending. In a sense, the distinction between borrowers and lenders is partly endogenous due to the fact that the distinction between efficient and inefficient entrepreneurs is endogenous. Similarly, AB endogenously determine the specialization of agents as borrowers or lenders. GZ endogenously determine the (long run or stationary) distribution between skilled and unskilled agents.

There is a strong correlation between the degree of capital market imperfections, the relevance of heterogeneity for macroeconomic performance, the relative complexity of the dynamics of the state variables involved in the analysis and the response of individuals and the aggregate to impulses.

Generally speaking, in the models surveyed so far, we can distinguish between the general case of *imperfect capital markets* – imperfections being generally *informational* in nature – and the special, simpler case of *perfect capital markets.* For instance, if we relax the assumption of equity rationing and bankruptcy costs in the GS model we go back to a perfect capital markets world in which the financial conditions of the individual firm do not matter as far as employment and production decisions are concerned. If we relax the assumption of monitoring cost in the BG model we go back to a perfect capital markets world in which the financial conditions of the individual entrepreneurs do not matter as far as capital accumulation is concerned.

In all these cases, the representative agent assumption is a legitimate working hypothesis. If employment and production decisions in the GS world or capital accumulation decisions in the BG world do not depend on individual financial conditions, but only on characteristics of the economy (for instance: technology) which are uniform across firms, heterogeneity is irrelevant and the *Modigliani–Miller* rule is reinstated. This is the *first best* case. It turns out that in this case the dynamics of the state variables under examination are relatively simple to analyze and determine a stable steady state (this is the GS' first best case) or the dynamics are trivial because . . . there are no dynamics, i.e. no feedback of the value of the state variable in one period to the values the variable will assume in the future (this is BGs first best case, for instance).

A slightly different story holds true for the KM framework, in which farmers and gatherers are differentiated not only because of the presence or absence of a financial constraint, but also because different agents can access different technologies. In particular farmers produce by means of a constant marginal returns technology while gatherers are characterized by a well behaved production function with decreasing marginal returns. This second distinction is inessential as to the deep characterization of the two groups of agents. In other words, what is really important in the model is the financial distinction and not the technological one. There is no reason, moreover, to link the nature of financially constrained (unconstrained) agent to the type of technology available.

The technological distinction, however, enables us to define in a peculiar and interesting way the perfect capital markets case. The special case of perfect capital markets in KM framework is obtained removing the crucial assumption of financial constraint but not the technological one. The first best is achieved when the marginal product of the durable non-reproducible asset (land) for the farmer, which is a given constant thanks to the constant marginal returns assumption, is equal to the marginal product of land for the gatherer, which is decreasing with the amount of land used in production by the gatherer thanks to the decreasing marginal returns assumption. This is tantamount to assuming a *no-arbitrage* condition.

Thanks to the technological differentiation, from the no-arbitrage condition KM can derive the amount of land allocated to each group of agents and the equilibrium values of the other variables without recurring to any laws of motion. In other words, in KMs framework, the first best case is characterized by (technological) heterogeneity but there are no dynamics. As we said before, the technological heterogeneity which persists in KM even when the financial constraints are removed is inessential as far as the macroeconomic implications of capital market imperfections are considered but they are necessary to derive univocally the allocation of the durable assets. If technology were uniform across agents the no-arbitrage condition would have left the allocation of land to farmers and gatherers undetermined.

Summing up, the perfect capital markets – Modigliani–Miller – first best is a special case characterized by the absence or irrelevance of heterogeneity (in our case, heterogeneity of financial conditions) which makes the representative agent assumption legitimate and the dynamics simple or even trivial. In this case shocks typically have symmetric effects. This is the consequence of homogeneity of employment, production and/or capital accumulation responses to exogenous shocks.

Symmetrically, in the imperfect capital markets case heterogeneity (in our case, heterogeneity of financial conditions) is essential, it invalidates the adoption of the representative agent assumption and generally determines a more interesting, sometimes complex dynamics. In this case shocks typically have asymmetric effects. This is the consequence of heterogeneity of employment, production and/or capital accumulation responses to exogenous shocks.

We now turn to aggregation. We can distinguish two procedures in the models discussed so far. In GS there is an initial continuous distribution of agents according to their level of equity base or net worth. This distribution can be characterized by its moments and in particular by the mean – the average equity base or the equity base of the average agent – and the variance. GS derive the law of motion of the average equity base, neglecting the role of the variance and its evolution over time. In the extension considered in Section 10.7, on the other hand, we show how to describe simultaneously the laws of motion of the average equity ratio and its variance. The initial distribution of the equity ratio may converge or not to a stationary distribution. As a matter of fact a wide range of dynamical behaviors are possible, within this framework depending upon the values assumed by the parameters.

The rest of the papers considered follows a different procedure. BG start from a continuous initial distribution of degrees of inefficiency and boil it down to a polarized distribution of efficient and inefficient entrepreneurs by means of a simple criterion of choice. GZ start from a continuous initial distribution of wealth and boil it down to a polarized distribution of skilled/rich individuals and unskilled/poor individuals by means of a piecewise linear phase diagram. Aggregate outcomes, therefore, emerge as weighted averages of polarized outcomes. In other words it is as if the modeler could compute the expected value of a discrete distribution consisting of only two characters. AB follow a similar line of inquiry in a more sophisticated setting. In fact, they treat the evolution over time of the distribution of wealth by means of the theory of convergence for monotonic Markov processes.

## Notes

1 Stoker (1993: 1829) notes that representative agent modelling ‘has proved a tremendous engine for the development of rational choice models over the last two decades, and their empirical application has developed into an ideology for judging aggregate data models’.

2 Kirman, 1992, is the *locus classicus* of this literature.

3 See, among others, Leontief, 1947; Gorman, 1953; Theil, 1954; Eisenberg, 1961; Fisher, 1969, 1982; Muellbauer, 1975, 1976; Lau, 1977, 1982. On aggregation see Daal and Merkies, 1984.

4 Lippi, 1988; Lippi and Forni, 1996; Lewbel, 1992.

5 Stoker, 1986,1993; Caballero et al., 1997; Geweke, 1985, demonstrate the pitfalls of using the representative agent framework to explain aggregate behaviour. Grunfeld and Griliches, 1960, show that macromodels with large aggregation bias can produce better prediction than their disaggregated counterparts. Empirical evidence on distribution of income and wealth in the growth process can be found in Kuznets, 1955, and Levy and Murnane, 1992.

6 Banerjee and Newman, 1988; Acemoglu, 1997; Benabou, 1993, 1996; Bertola, 1993; Hopenhayn, 1992; Solon, 1992; Persson and Tabellini, 1994; Zarembka, 1975; Chiu, 1998.

7 Roller and Sinclair-Desgagne, 1996; Bertola and Caballero, 1994; Keane et al., 1988; Heckman and Sedlacek, 1985.

8 Keane, 1997; Amable and Chatelain, 1997; Guariglia and Schiantarelli, 1998; Zietz, 1996; Fortin, 1995; Caballero, 1995; Gordon, 1992; Perotti, 1993; Attanasio and Weber, 1995.

9 Cho, 1995; Aoki, 1994; Chiaromonte and Dosi, 1993; Caballero and Engel, 1993; Das, 1993; Zeira, 1994; Caplin and Leahy, 1991, 1994; Kydland, 1984.

10 Aghion and Bolton, 1997; Banerjee and Newman, 1988, 1993; Galor and Zeira, 1993; Piketty, 1997; Bernanke and Gertler, 1989, 1990; Greenwald and Stiglitz, 1988, 1990, 1993; Kiyotaki and Moore, 1997.

11 In other words, there are marginal decreasing returns. The analysis could be carried out, however, also in the presence of constant marginal returns, i.e. in the case in which $F$ is linear.

12 When capital markets are affected by informational imperfections such as asymmetric information, a *financing hierarchy (pecking order)* can be envisaged. Internal finance is the most preferred source of finance. As to external sources, credit has a cost advantage over the issue of new equities (Fazzari, Hubbard and Petersen, 1988). A different wording is used in the literature on the so-called *credit view:* bank loans are imperfect substitutes of the issue of new equities (credit is ‘special’). See, for instance, Bernanke and Blinder (1988).

13 As a matter of fact, the procedure followed by GS to derive an equation for the probability of bankruptcy is much more complicated. In this survey we can simplify the argument without loss of generality.

14 On bankruptcy costs, see Gordon and Malkiel, 1981; Altman,1984; Gilson, 1990; Kaplan and Reishus, 1990.

15 There is only demand for consumption goods in this framework. Investment is ruled out by assumption, since production takes place using only labour as an input. Financial factors do not play any role in the determination of aggregate demand, due to the peculiar way in which consumer’s preferences and the budget constraint are modelled. It is clear, however, that if firms use capital together with labour to carry on production, financial factors can influence aggregate demand through their impact on investment activity. Even in the absence of investment activity, however, demand can be influenced by financial factors. If firms pay-out dividends based on their net worth, in fact, and dividends are part of consumers’ income, consumption expenditure is affected by firms’ financial conditions.

A different and somehow more indirect way to explore the impact of financial factors on aggregate demand is adopted in Delli Gatti and Gallegati (1997).

16 This point is discussed at length in Delli Gatti and Gallegati (1997).

17 BG do not impose equilibrium on the market for loanable funds. They simply assume that the volume of loans available is sufficient to fill the financing gap of borrowers. The rate of return on storage therefore is not determined in equilibrium but is exogenous.

18 For the sake of simplicity, we are ignoring the effects on the price and the quantity of capital of changes of other exogenous variables, such as the interest rate and the share of entrepreneurs in total population.

19 For the sake of simplicity, we are ignoring the effects on the price and the quantity of capital of changes of other exogenous variables, such as the interest rate, the share of entrepreneurs in total population, the cost of monitoring etc.

20 On this issue see Hart and Moore (1994, 1998).

21 As a matter of fact, the paper by Aghion and Bolton has been circulating for some years as a LSE working paper.

22 As a matter of fact, AB show that under appropriate conditions there can be credit rationing. We will not dig deeper into this issue.

23 If agents were identical – i.e. if the Representative Agent Hypothesis held true – the law of motion of the equity ratio of the representative firm would be:

$$a_t = \Gamma_1 a_{t-1} - \Gamma_2 a_{t-1}^2 + \Gamma_0$$

which is a quadratic map topologically conjugated to the logistic map.
The dynamical properties of the logistic map are well known (for a comprehensive survey see Day, 1994). It is worth mentioning, however, that there are configurations of the parameters such that the dynamics are chaotic, i.e. the equity ratio oscillates apparently at random around the steady state. In this case the economy follows a path of endogenously determined fluctuating growth.

24 See Agliari et al. (2000) for a thorough analysis of the dynamic properties of this type of system.

## References and further reading

Aaron, H. and Pechman, J. (eds) (1981) *How Taxes Affect Economic Behaviour*. Washington, D.C.: Brookings Institution.

Abramovitz, M. (ed.) (1959) *The Allocation of Economic Resources*, Stanford, California: Stanford University Press

Acemoglu, D. (1997) ‘Matching, Heterogeneity, and the Evolution of Income Distribution’, *Journal of Economic Growth*, 2, pp. 61–92.

Aghion, P. and Bolton, P. (1997)‘A Trickle Down Theory of Growth and Distribution with Debt Overhang’, *Review of Economic Studies* , 64, pp. 151–72.

Agliari, A., Delli Gatti, D., Gallegati, M. and Gardini, L. (2000) ‘Global Dynamics in a Non-linear Model of the Equity Ratio’, em Journal of Chaos, Solitons and Fractals., 11, pp. 961–85.

Akerlof, G. (1970)‘The Market for ‘Lemons’: Quality Uncertainty and the Market Mechanism’, *Quarterly Journal of Economics*, 85, pp. 488–500.

Altman, E. (1984) ‘A Further Empirical Investigation of the Bankruptcy Cost Question’, *Journal of Finance*, 39, pp. 1067–89.

Amable, B. and Chatelain, J.B. (1997) ‘How Do Public Investment and Financial Factors Affect Growth in a Debt Overhang Economy?’, *Manchester School of Economic and Social Studies*, 65, pp. 310–27.

Anderlini, L. and Ianni, A. (1996) ‘Path dependence and learning from neighbours’, *Games and Economic behaviour, 13*, pp.141–77.

Anderson, P.W., Arrow, K.J. and Pines, D.(eds) (1988) *The Economy as a Complex Evolving System*, Redwood City: Addison-Wesley.

Aoki, M. (1994) ‘New Macroeconomic Modeling Approaches: Hierarchical Dynamics and Mean Field Approximation’, *Journal of Economic Dynamics and Control*, 18, pp. 865–77.

Aoki, M. (1995) ‘Economic Fluctuations with Interactive Agents: Dynamic and Stochastic Externalities’, *Japanese Economic Review*, 46, pp.148–65.

Aoki, M. (1997) *New approaches to macroeconomics modeling: evolutionary stochastic dynamics, multiple equilibria, and externalities as field effects*, New York: Cambridge University Press.

Arrow, K.J. (1951) *Social Choice and Individual Values*, New Haven, Yale University Press.

Arrow, K.J. (1959) ‘Towards a Theory of Price Adjustment’ in M. Abramovitz (ed.), pp. 41–51.

Attanasio, O. and Weber, G. (1995) ‘On the Aggregation of Euler Equations for Consumption in Simple Overlapping-Generation Models’, *Economica*, 62, pp. 565–76.

Banerjee, A. and Newman, A. (1988) ‘Risk-Bearing and the Theory of Income Distribution’, *Review of Economic Studies*, 58, pp. 211–36.

Banerjee, A. and Newman, A. (1993) ‘Occupational Choice and the Process of Development’, *Journal of Political Economy*, 101, pp.274–88.

Becker G.S. (1962) ‘Irrational behaviour and Economic Theory’, *Journal of Political Economy*, 70, pp. 1–13.

Benabou, R. (1993) ‘Search Market Equilibrium, Bilateral Heterogeneity, and Repeated Purchases’, *Journal of Economic Theory*, 60, pp. 140–58.

Benabou R. (1996) ‘Heterogeneity, Stratification and Growth’, *American Economic Review*, 86, pp. 584–609.

Bernanke, B. and Blinder, A. (1988) ‘Credit, Money, and Aggregate Demand’, *American Economic Review, Papers and Proceedings*, 78, pp. 435–39.

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

Bernanke, B. and Gertler, M. (1990) ‘Financial Fragility and Economic Performance’, *Quarterly Journal of Economics*, 105, pp. 87–114.

Bertola, G. (1993) ‘Factor Shares and Savings in Endogenous Growth’, *American Economic Review*, 83, pp. 1184–98.

Bertola, G. and Caballero, R.J. (1994) ‘Irreversibility and Aggregate Investment’, *Review of Economic Studies*, 61, pp. 223–46.

Blume, L. (1993) ‘The statistical mechanics of strategic interactions’, *Games and Economic behaviour*, 5, pp. 387–424.

Brock, W.A., and Durlauf, S.N. (1995) ‘Discrete Choice with Social Interactions I: Theory’, *NBER working paper 5291*.

Brunner, K. and Meltzer, A. (eds), (1977) *Stabilization of the Domestic and International Economy*. Carnegie-Rochester Conference Series on Public Policy, 5, Amsterdam: North-Holland.

Caballero, R. J. (1992) ‘A Fallacy of Composition’, *American Economic Review*, 82, pp. 1279–92.

Caballero, R. J. (1995) ‘Near Rationality, Heterogeneity, and Aggregate Consumption’, *Journal of Money, Credit, and Banking*, 27, pp. 29–48.

Caballero, R. J. and Engel, E. (1993) ‘Microeconomic Adjustment Hazards and Aggregate Dynamics’, *Quarterly Journal of Economics*, 108, pp. 359–83.

Caballero, R.J., Engel, E. and Haltinwanger, J. (1997) ‘Aggregate Employment Dynamics: Building from Microeconomic Evidence’, *American Economic Review*, 87, pp. 115–37.

Caplin, A. and Leahy, J. (1991) ‘Sectoral Shock, Learning, and Aggregate Fluctuations’, *Review of Economic Studies*, 60, pp. 777–94.

Caplin, A. and Leahy, J. (1994) ‘Business as Usual, Market Crashes, and Wisdom after the Fact’, *American Economic Review*, 84, pp. 548–65.

Chiaromonte, F. and Dosi, G. (1993) ‘Heterogeneity, Competition, and Macroeconomic Dynamics’, *Structural Change and Economic Dynamics*, 4, pp. 39–63.

Chiu, W.H. (1998) ‘Income Inequality, Human Capital Accumulation and Economic Performance’, *Economic Journal*, 108, pp. 44–59.

Cho, J. (1995) ‘Ex post Heterogeneity and the Business Cycle’, *Journal of Economic Dynamics and Control*, 19, pp. 533–51.

Colander, D. (1994) 'The Macrofoundations of Microeconomics', *Eastern Economic Journal*, 19, pp. 447–58.

Colander, D. (ed.) (1996) *Beyond Microfoundations: Post Walrasian Macroeconomics*, Cambridge: Cambridge University Press.

Cooley, T. F. (ed.) (1995) *Frontiers of Business Cycle Research*, Princeton, New Jersey: Princeton University Press.

Daal, J. and Merkies, A. (1984) *Aggregation in Economic Research: From Individual to Macro Relations*, Dordrecht: D. Reidel.

Das, S. P. (1993) *New perspectives on business cycles: An analysis of inequality and heterogeneity*, Aldershot, U.K.: Elgar.

Day, R. H. (1994) *Complex Economic Dynamics*, Cambridge, MA: MIT Press.

Delli Gatti, D. and Gallegati, M. (1995) 'Financial Fragility and Economic Fluctuations: Keynesian Views (and an addendum)', *Economic Notes*, 24, pp. 513–54.

Delli Gatti, D. and Gallegati, M. (1997) 'Financial Constraints, Aggregate Supply and the Monetary Transmission Mechanism' *Manchester School of Economic and Social Studies*, 65, pp. 101–26.

Delli Gatti, D., Gallegati, M. and Kirman, A. (eds) (2000) *Interaction and Market Structure*, Berlin: Springer.

Delli Gatti, D., Gallegati, M. and Palestrini, P. (2000) 'Agents' Heterogeneity, Aggregation and Economic Fluctuations', in Delli Gatti, D., Gallegate, M. and Kirman, A. (eds)

Eisenberg, B. (1961) 'Aggregation of Utility Functions', *Management Science*, 6, pp. 337–350.

Ellison, G. (1993) 'Learning, Local Interaction and Coordination', *Economica*, 61, pp. 1047–72.

Fazzari, S., Hubbard, G. and Petersen, B. (1988) 'Financing Constraints and Corporate Investment', *Brookings Papers on Economic Activity*, 1, pp. 141–206.

Fisher, F.M. (1969) 'The Existence of Production Functions', *Econometrica*, 37, pp. 553–77.

Fisher F.M. (1982) 'Aggregate Production Functions Revisited', *Review of Economics and Statistics*, 49, pp. 615–26.

Follmer, H. and Majumdar, M. (1978) 'On the Asymptotic Behaviour of Stochastic Economic Processes. Two Examples from Intertemporal Allocation under Uncertainty', *Journal of Mathematical Economics*, 5, pp. 275–87.

Fortin, N. (1995) 'Heterogeneity Biases, Distributional Effects, and Aggregate Consumption', *Journal of Applied Econometrics*, 10, pp. 287–311.

Gaffeo, E. (1999) 'Tutorial in Social Interaction Economics', in Gallegati, M. and Kirman, A. (eds), pp. 47–76

Gallegati, M. and Kirman, A. (eds) (1999) *Beyond the Representative Agent*, Aldershot, UK: Elgar.

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

Gertler, M. (1988) 'Financial Structure and Aggregate Economic Activity', *Journal of Money, Credit and Banking*, 20, pp. 559–88.

Geweke, J. (1985) 'Macroeconometric Modeling and the Theory of the Representative Agent', *American Economic Review*, 75, pp. 206–10.

Gilson, S. (1990) 'Bankruptcy, Boards, Banks and Blockholders: Evidence on Changes in Corporate Ownership and Control when Firms Default', *Journal of Financial Economics*, 27, pp. 355–88.

Gordon, S. (1992) ‘Costs of Adjustment, the Aggregation Problem and Investment’, *Review of Economics and Statistics*, 74, pp. 422–29.

Gordon, R. and Malkiel, B. (1981) ‘Corporation Finance’, in Aaron, H. and Pechman, J. (eds)

Gorman, W. M. (1953) ‘Community Preference Fields’, *Econometrica*, 21, pp. 63–80.

Grandmont, J.M. (1985) ‘On Endogenous Competitive Business Cycles’, *Econometrica*, 53, pp. 995–1045.

Green, H. (1964) *Aggregation in Economic Analysis: An Introductory Survey*, Princeton, New Jersey: Princeton University Press.

Green, H. (1977) ‘Aggregation Problems of Macroeconomics’, in G. C. Harcourt (ed.), pp. 179–94.

Greenwald, B. and Stiglitz, J. (1986) ‘Externalities in Economies with Imperfect Information and Incomplete Markets’, *Quarterly Journal of Economics*, 101, pp. 229–64.

Greenwald, B. and Stiglitz, J. (1988) ‘Imperfect Information, Finance Coinstraints and Business Fluctuations’, in Kohn, M. and Tsiang, S. (eds)

Greenwald, B. and Stiglitz, J. (1990) ‘Macroeconomic Models with Equity and Credit Rationing’, in Hubbard, R.G. (ed.)

Greenwald, B. and Stiglitz, J. (1993) ‘Financial Market Imperfections and Business Cycles’, *Quarterly Journal Of Economics*, 108, pp. 77–114.

Greenwald, B., Stiglitz, J. and Weiss, A. (1984) ‘Informational Imperfections in the Capital Markets and Macroeconomic Fluctuations’, *American Economic Review*, 74, pp. 194–9.

Grunfeld, Y. and Griliches Z. (1960) ‘Is Aggregation Necessarily Bad?’ *Review of Economics and Statistics*, 42, pp. 1–13.

Guariglia, A. and Schiantarelli, F. (1998) ‘Production smoothing, firms heterogeneity, and financial constraints’, *Oxford Economic Papers*, 88, pp. 63–78.

Haan, W. (1997) ‘Solving Dynamic Models with Aggregate Shocks and Heterogeneous Agents’, *Macroeconomic Dynamics*, 1, pp. 355–86.

Hahn, F. H. (1973) *On the Notion of Equilibrium in Economics*. Cambridge, UK: Cambridge University Press.

Hahn, F.H. (1984) *Equilibrium and Macroeconomics*, Oxford: Basil Blackwell.

Hahn, F.H. and Solow, R. (1995) *A Critical Essay on Modern Macroeconomic Theory*, Cambridge, Massachusetts: MIT Press.

Harcourt, G.C. (ed.) (1977) *The Microeconomic Foundations of Macroeconomics*, Boulder, Colorado: Westview Press.

Hardle, W. and Kirman, A.(1995) ‘Non Classical Demand. A Model-Free Examination of Price-Quantity Relations in the Marseille Fish Market’, *Journal of Econometrics*, 67, pp. 227–57.

Hart, O. and Moore, J. (1994) ‘A Theory of Debt Based on the Inalienability of Human Capital’, *Quarterly Journal of Economics*, 109, pp. 841–79.

Hart, O. and Moore, J. (1998) ‘Default and Renegotiation: a Dynamic Model of Debt’, *Quarterly Journal of Economics*, 1113, pp. 1–41.

Hartley, J. (1997) *The Representative Agent in Macroeconomics*, London: Routledge.

Heckman, J.J. and Sedlacek, G. (1985) ‘Heterogeneity, Aggregation, and the Market Wage Functions’, *Journal of Political Economy*, 93, pp. 1077–125.

Hopenhayn, H. (1992) ‘Entry, Exit, and Firm Dynamics in Long Run Equilibrium’, *Econometrica*, 60, pp. 1127–50.

Hubbard, R.G. (1990) *Asymmetric Information, Corporate Finance and Investment*, Chicago: University of Chicago Press.

Kaplan, S. and Reishus, D. (1990) 'Outside Directorship and Corporate Performance', *Journal of Financial Economics*, 27, pp. 389–410.

Keane M., Moffit, R. and Runkle, D. (1988) 'Real Wages over the Business Cycle', *Journal of Political Economy*, 96, pp. 1232–66.

Keane, M. (1997) 'Modeling Heterogeneity and State Dependence in Consumer Choice behaviour', *Journal of Business and Economic Statistics*, 15, pp. 310–27.

Kirman, A. (1992) 'Whom or What Does the Representative Individual Represent?' *Journal of Economic Perspectives*, 6, pp. 117–36.

Kirman, A. (1993) 'Ants, Rationality, and Recruitment', *Quarterly Journal of Economics*, 107, pp. 136–57.

Kirman, A. (1999) 'Interaction and Markets', in Gallegati, M. and Kirman, A. (eds).

Kiyotaki, N. and Moore, J. (1997) 'Credit Cycles', *Journal of Political Economy*, 105, pp. 211–48.

Kohn, M. and Tsiang, S. (eds) (1988) *Finance Constraints, Expectations and Macroeconomics*, Oxford: Oxford University Press.

Kuznets, S. (1955) 'Economic Growth and Income Inequality', *American Economic Review*, 45, pp. 1–28.

Kydland, F. (1984) 'Heterogeneous Agents in Quantitative Aggregate Economic Theory', *Journal of Economic Dynamics and Control*, 18, pp. 849–64.

Kydland, F. and Prescott, E. (1982) 'Time to Build and Aggregate Fluctuations', *Econometrica*, 50, pp. 1345–70.

Lau, L.J. (1977) 'Existence Conditions for Aggregate Demand Functions', *Institute for Mathematical Studies in the Social Science, Stanford University, Technical Report 248*.

Lau, L.J. (1982) 'A Note on the Fundamental Theorem of Exact Aggregation', *Economics Letters*, 9, pp. 119–26.

Leijonhuvfud, A. (1981) *Information and Coordination*, New York: Oxford University Press.

Leijonhuvfud, A. (1992) 'Keynesian Economics: Past Confusions, Future Prospects', in Vercelli, A. and Dimitri, N. (eds), pp. 16–37.

Leijonhufvud, A. (1993) 'Towards a not-too-rational macroeconomics', *Southern Economic Journal*, 60, pp. 1–13.

Leijonhuvfud, A. (1995) 'Adaptive behaviour, market processes and the computable approach', *CCE w.p. #20*.

Leontief, W. (1947) 'Introduction to a Theory of the Internal Structure of Functional Relationships', *Econometrica*, 15, pp. 361–73.

Levy, F. and Murnane, R. (1992) 'US Earnings Levels and Earnings Inequality', *Journal of Economic Literature*, 30, pp. 1333–81.

Lewbel, A. (1992) 'Aggregation and Simple Dynamics', *mimeo*, Brandeis University.

Lippi, M. (1988) 'On the Dynamic Shape of Aggregated Error Correction Models', *Journal of Economic Dynamics and Control*, 12, pp. 561–85.

Lippi, M. and Forni, M. (1996) *Aggregation and the Microundations of Macrodynamics*, Oxford: Oxford University Press.

Long, J.B. and Plosser, R. (1983) 'Real Business Cycle', *Journal of Political Economy*, 91, pp. 39–69.

Lucas, R. (1975) 'An Equilibrium Model of the Business Cycle', *Journal of Political Economy*, 83, pp. 1113–44.

Lucas, R. (1977) 'Understanding Business Cycles,' in Brunner, K. and Meltzer, A. (eds) pp. 7–29.

Mankiw, N. and Romer, D. (eds) (1991) *New Keynesian Economics*, 2 vols., Cambridge, Massachusetts: MIT Press.

Mantel, R. (1976) 'Homothetic Preferences and Community Excess Demand Functions', *Journal of Economic Theory*, 12, pp. 197–201.

Marshall, A. (1920) *Principles of Economics*, 9th variorum ed., London: Macmillan.

Martel, R. (1996) 'Heterogeneity, Aggregation and a Meaningful Macroeconomics', in Colander, D. (ed.), pp. 127–44.

Minsky, H.P. (1982) *Can 'It' Happen Again?* New York: M.E. Sharpe.

Morris, S. (1997) 'Interaction games: a unified analysis of incomplete information, local interaction and random matching', *mimeo*.

Muellbauer, J. (1975) 'Aggregation, Income Distribution and Consumer Demand', *Review of Economic Studies*, 42, pp. 525–44.

Muellbauer, J. (1976) 'Community Preferences and the Representative Consumer', *Econometrica*, 44, pp. 979–1000.

Myers, S. and Majluf, N. (1984) 'Corporate Financing and Investment Decisions when Firms have Information that Investors do not have', *Journal of Financial Economics*, 13, pp. 187–221.

Perotti, R. (1993) 'Income Distribution and Investment', *European Economic Review*, 38, pp. 827–35.

Persson, T. and Tabellini, G. (1994) 'Is Inequality Harmful for Growth? Theory and Evidence', *American Economic Review*, 48, pp. 600–21.

Piketty, T. (1997) 'The Dynamics of the Wealth Distribution and the Interest Rate with Credit-Rationing', *Review of Economic Studies*, 64, pp. 173–89.

Quah, J. (1997) 'The Law of Demand When Income Is Price Dependent', *Econometrica*, 65, pp. 1421–42.

Rios-Rull, J.V. (1995) 'Models with Heterogenous Agents', in Cooley, T.F. (ed.), pp. 98–125.

Rizvi, S. (1994) 'The Microfoundations Project in General Equilibrium Theory', *Cambridge Journal of Economics*, 18, pp. 357–77.

Roller, L. and Sinclair Desgagne, B. (1996) 'On the Heterogeneity of Firms', *European Economic Review*, 40, pp. 531–9.

Sargent, T. (1993) *Bounded Rationality in Macroeconomics*, Oxford: Clarendon Press.

Schumpeter, J. A. (1951) The Historical Approach to the Analysis of Business Cycles, *NBER Conference on Business Cycles*.

Schumpeter, J.A. (1954) *History of Economic Analysis*, Oxford, Oxford University Press.

Selten, R. (1964) 'Valutation of N-person Games', in *Advances in Game Theory*, 52, pp. 565–78.

Solon, G. (1992) 'Intergenerational Income Mobility in the United States', *American Economic Review*, 102, pp. 393–408.

Sonnenschein, H. (1972) 'Market Excess Demand Functions', *Econometrica*, 40, pp. 549–63.

Stiglitz, J. (1969) 'Distribution of Income and Wealth among Individuals', *Econometrica*, 37, pp. 382–97.

Stiglitz, J. (1992) 'Methodological Issues and the New Keynesian Economics', in Vercelli, A. and Dimitri, N. (eds), pp. 38–86.

Stoker, T. (1986) 'Simple Tests of Distributional Effects on Macroeconomic Equations', *Journal of Political Economy*, 94, pp. 763–95.

Stoker, T. (1993) 'Empirical Approaches to the Problem of Aggregation over Individuals', *Journal of Economic Literature*, 21, pp. 1827–74.

Theil, H. (1954) *Linear Aggregation of Economic Relations*, Amsterdam: North Holland.

Vercelli, A. and Dimitri, N. (eds) (1992) *Macroeconomics: A Survey of Research Strategies*, Oxford: Oxford University Press.

Weiss, Paul (1967) ‘1+1≠2: (One Plus One Does Not Equal Two)’ in Garder C., Quarton, T. and Schmitt, O. (eds) pp. 801–21.

Zarembka, P. (1975) ‘Capital Heterogeneity, Aggregation, and the Two-Sector Model’, *Quarterly Journal of Economics*, 89, pp. 103–14.

Zeira, J. (1994) ‘Informational Cycles’, *Review of Economic Studies*, 61, pp. 31–44

Zietz, J. (1996) ‘Aggregate Consumption with Heterogeneous Agents and a Changing Income Distribution’, *Atlantic Economic Journal*, 24, pp. 361–70.

# 11 Toward the microeconomics of innovation

## Growth engine of market economies

*William J. Baumol*

> The Bourgeoisie (i.e. capitalism cannot exist without constantly revolutionizing the instruments of production. . . . Conservation of the old modes of production in unaltered form was, on the contrary, the first condition of existence for all earlier industrial classes. . . . The bourgeoisie, during its rule of scarce one hundred years has created more massive and more colossal productive forces than have all preceding generations together. . . . It has accomplished wonders far surpassing Egyptian pyramids, Roman aqueducts and Gothic cathedrals. . . .
>
> (Marx and Engels, *The Communist Manifesto*, 1847)

### 11.1 Introduction: the *big* puzzle: why do all rival systems trail so far behind free market growth rates?

Undoubtedly, the spectacular and unmatched growth rate of the industrialized free-market economies is what most distinguishes them from *all* other economic systems. In no other system, current or in the past, has the average income of the general public risen anywhere nearly as much or as quickly as it has in North America, Western Europe and Japan. Though the former Soviet Union planned its economy and forced its population to invest heavily in factories and hydroelectric dams, its failure to produce enough to raise the standard of living of its population to that of the free-market economies undoubtedly played a major role in its downfall. There have been great civilizations with extraordinary records of invention and engineering – medieval China and ancient Rome are clear examples. But none has approached the growth record of modern free-market economies. What is the secret of their extraordinary success? That is the economic puzzle that is absolutely critical to the degree of prosperity our future is able to achieve. Its solution is what the world's poorer countries are anxious to learn.

## 11.2 The free market's growth record

The growth of per-capita income and productivity in the free-market economies is so enormous that it is virtually impossible to comprehend. In contrast, average growth rates of per-capita incomes were probably approximately zero for about 1,500 years before the Industrial Revolution at the time of George Washington. In 1776, even the wealthiest consumers in England, then the world's richest country, had only a half-dozen consumer goods that had not been present in ancient Rome. These new products included (highly inaccurate) hunting guns, (highly inaccurate) watches, paper, window glass and very little else. And, remarkably enough, Roman citizens enjoyed a number of amenities, such as hot baths and good roads, that had long disappeared at the time of the American Revolution.

In contrast, in the past century and a half, per-capita incomes in the typical capitalist economy have risen by amounts ranging from several hundred to several thousand percent. Recent decades have yielded an unmatched outpouring of new products and services: color television, the computer, jet aircraft, the VCR, the microwave oven, the hand-held calculator, the cellular telephone, and so on and on. And the flood of new products continues. When, a few years ago, many of the world's communist regimes collapsed and when even the masters of China turned toward capitalist enterprise, surely part of the reason was the public's desire to participate in the growth miracle of the capitalist economies that Marx and Engels – those high priests of anticapitalist movements – were among the first economists to discern (as the opening quotation demonstrates).

The explanation of this miracle must be sought in the activities of industries and the business firms of which they are constituted, for they are the producers of the increasing outpouring of goods and services that constitutes the growth record of capitalism. It must be something about business firms and the decisions they make that plays a vital part in this prime attribute of our economy. Yet the standard core microeconomic theory of firms and industries, while it has included some outstanding contributions on the theory of innovation, has not provided anything suggesting what features of business behavior and decision-making can account for all this growth. Indeed, as we know, mainstream microeconomics offers reasons to expect that the capitalist economy will be characterized by a growth performance that is far from optimal.

Here, I will describe some features of competitive markets to which the growth performance of business firms can be attributed, features that literally force businesses to do all they can to contribute to the growth miracle. I will then provide some hints of a microeconomic model, using the most elementary of microeconomic tools to analyze this process.

## 11.3 Innovation and the growth process

Without a doubt, a primary source of the growth miracle of the past two centuries is 'the wave of new gadgets'[1] – the surge of innovation that probably first reached a substantial pace in the first third of the nineteenth century. Though it is difficult to prove statistically (because improved education and the construction of factories, roads and other influences certainly made substantial contributions), a very large proportion of all of the economic growth that has occurred since the eighteenth century probably is ultimately attributable to innovation. Indeed, the incredible poverty of earlier centuries that the inventions of the Industrial Revolution helped to bring to an end meant that previously society simply could not afford to spend much on education or on the construction of plant and machinery. For us, the magnitude of that poverty is difficult to grasp. The following passage from the writings of a noted historian suggests how serious the problem was only a few centuries ago:

> The poor in the towns and countryside lived in a state of almost complete deprivation. Their furniture consisted of next to nothing. . . . Inventories made after death . . . testify almost invariably to the general destitution . . . a few old clothes, a stool, a table, a bench, the planks of a bed, sacks filled with straw. Official reports for Burgundy between the sixteenth and the eighteenth centuries are full of references to people (sleeping) on straw . . . with no bed or furniture, who were only separated from the pigs by a screen. . . . Paradoxically the countryside sometimes experienced far greater suffering (from famines than the townspeople). The peasants . . . had scarcely any reserves of their own. They had no solution in case of famine except to turn to the town where they crowded together, begging in the streets and often dying in public squares. . . .
>
> (Fernand Braudel, *The Structures of Everyday Life*, Vol. I, New York: Harper and Row, 1979, pp. 73–75 and 284–286).

Only the growing outputs that innovation, first in agriculture and mining and then in manufacturing and transportation, made feasible produced the enormous increases in productive plant and equipment and in education (and other forms of investment in human capital) that are widely judged to have contributed greatly to economic growth. Thus, it can be argued not only that innovation has facilitated the growth process, but that without it the process would have been reduced to insignificance.

Two of the leading analysts of economic growth conclude:

> As yet, no empirical study proves that technology has been the engine of modern-day growth. Still, we ask the reader to ponder the following: What would the century's growth performance have been like without the invention and refinement of methods for generating electricity and using radio waves to transmit sound, without Bessemer's discovery of a new technique for refining iron, and without the design and development of products like the automobile, the airplane, the transistor, the integrated circuit, and the computer?
>
> (Gene M. Grossman and Elthanan Helpman, 'Endogenous Innovation in the Theory of Growth,' *Journal of Economic Perspectives*, Vol. 8, Winter, 1994, p. 32).

*Table 11.1* Some leading growth macromodels

| | | |
|---|---|---|
| *A. Autonomous innovation:* | | |
| 1) | $Y_t = A(t)f(L_t)$ | Ricardo (1817) |
| 2) | $Y = A(t)K^a L^{(1-a)}$ | Solow (1956) |
| *B. Endogenous innovation* | | |
| 3) | $Y_j = A(K)K_j^a L_j^{(1-a)}$ | Arrow (1962) (spillovers, learning by doing) |
| 4) | $Y_j = A(R)F(R_j, K_j, L_j)$ | Romer (1986) (R = Knowledge) |
| 5) | $Y_j = A(H)F(K_j, H_j)$ | Lucas (1988) (H = Human capital) |
| 6) | $Y = A(H)K^a H^b L^{(1-a-b)}$ | Mankiw, D. Romer and Weil (1992) |

## 11.4 Macroeconomic endogenous growth models: where is the Prince of Denmark?

The model building that is introduced here relies heavily on the current macroeconomic growth literature. None of what follows, consequently, is to be interpreted as criticism, much less denigration, of the earlier writings. However, since it is my hope to carry the study of the subject a step beyond what that work has been able to achieve, I must begin by indicating what this literature has not yet succeeded in doing. In my view left by that literature is its failure to grapple with the extraordinary growth record of the capitalist economies. Indeed, as is well known, and as my brief review of this literature will note, the earlier contributions took innovation to be an autonomous contribution of the passage of time. It was, in effect, described as a sort of manna dropped in a steady stream from some unspecified source, and that could just as well emerge from a capitalistic economy or from any other. Later model builders recognized that this formulation was inadequate, and that there were features inherent in the economic processes that account for innovation and growth. This led to the valuable line of analysis referred to as 'endogenous growth theory.' Yet the features cited in this literature as sources of innovation, notably the externalities of innovation, and the acquisition of human capital, in part through learning by doing, apply to many forms of economic organization, and not only to the free-market economies.

In my view, these very valuable contributions are like performances of *Hamlet* that include the King, Ophelia, Gertrude, and many other of the crucial characters, but omit the Prince of Denmark. They tell us much about innovation and growth, but they fail to account for the most salient and extraordinary feature of the growth record, the entirely unparalleled success of the free-market economies. I will suggest that they fail to do so because they are macromodels, something patently unobjectionable in itself, but that is a major handicap for study of the issue before us, which, I believe, is explainable primarily in terms of *microeconomic* behavior.

Recent growth analysis had its beginnings in the 1950s with the work of Solow and Swan – work which deservedly elicited renewed interest in models of growth and in approaches compatible with statistical estimation. The models themselves represented no break with the past, and clearly have their roots in the work of the classical economists, notably that of David Ricardo. Table 11.1 is a vastly oversimplified summation of some of the leading models in the more recent group, including also the Ricardian model for contrast.

The Ricardian model is sufficiently familiar and needs little review here. In short, it postulates linear relationships throughout, with the exception of diminishing returns to labor and capital invested on a given stock of land. In an early stage of the economy, with highly productive land abundant, output, in the model, is more than sufficient to provide subsistence to agricultural workers. This initially yields high profits, induces increased innovation and expands the demand for labor. Higher wages stimulate an expanded population, and diminishing returns mean that a second round in this parable yields a second set of increases in profits, wages, population and output, but all of them smaller than those in the previous round. This process continues until, finally, diminishing returns cut output to a level only capable of providing subsistence wages to workers, and there the process would end in the stationary state, were it not for innovation. But Ricardo and other classical economists recognized that innovation *does* occur. This results in a shifting of the production function and postponement of the stationary state, something that can occur repeatedly and can keep the economy expanding indefinitely. What is missing in the Ricardian story is any explanation of the innovation process, and certainly of any endogenous innovation model. That is why the innovation process is represented simply as $A(t)$, as a function of time and nothing else, and with no distinguishing features that differentiate the process in a capitalist economy from that in any other form of economic organization. Thus, Ricardo's story emphatically contains no role for the Prince of Denmark.

The original Solow model, the prototype neoclassical model, contains a representation of innovation not much different from Ricardo's, with innovation also autonomous, and undifferentiated as between free-market economies and other economic forms. The model also assumes that there are diminishing returns to capital, an attribute that can be used as a hypothesis that predicts convergence of productivities and per-capita incomes in different economies, because wealthier economies have relatively large capital stocks whose productivities, relative to those of poorer countries, are consequently reduced severely by diminishing returns.

Two observations led Romer to argue that the neoclassical model required some modification. First, he observed that the universal convergence apparently predicted by the theory is not sustained by the facts. Indeed, the many statistical studies of the convergence hypothesis generally conclude that while the wealthiest economies have, indeed, been converging toward one another's productivities and per-capita incomes, most of the impecunious nations are falling further behind. Second, he noted (as students of the subject such as Schmookler had long observed) that the innovation process is neither largely autonomous nor largely fortuitous. The amount of activity devoted to innovation, and the output of that activity, is influenced substantially by what is going on in the economy. This argument served to reorient research toward the endogenous growth models. Some of these, including several that preceded Romer's writings, are also characterized in Table 11.1. For example, the Lucas model of 1988 can be described as taking the innovation function as $A(H)$, where $H$ is the investment in human capital of the entire society, as distinguished from $H_j$, the corresponding investment by agent

$J$.[2] Similarly, the Arrow model of 1962 uses an innovation function such as $A(K)$, where $K$ is society's investment in physical capital, as distinguished from that of agent $J$. The other entries in Table 11.1 can easily be interpreted analogously by the reader.

Again, what is to be noted is that none of these formulations distinguish the free-market economy from other economic forms. Thus, whatever their virtues, none of them assigns a part in the scenario to the Prince of Denmark. Nor should this be surprising. Other economies, both historical and modern, have stressed education, have innovated, have experienced spillovers from innovation, in short, they have exhibited all the endogenous innovation features of the newer models. None of these new models seems to have stressed any special features of the capitalist economy. To do so, I believe, it is necessary to turn to microeconomics.

## 11.5 What is different about free-market economies?

Invention alone is not the complete answer to the great puzzle – the explanation of the free market's unmatched growth performance. Earlier societies have had a spectacular invention record. The Chinese are the outstanding example. Centuries before Columbus they had invented printing, the compass, complex clockwork, gunpowder, spinning machinery, a cotton gin, porcelain, matches, toothbrushes, playing cards and much more. There have been other countries in history with a considerable record of new products and new technology. Moreover, education was highly valued in the Chinese culture and others, though, it must be admitted, much of the population was uneducated. Yet, economically, these inventions and this education never produced economic growth anything like that in the modern market economies.[3]

It should be added that markets of substantial importance exist in virtually every economy of the world and have existed throughout recorded history. What, then, is different about modern markets that not only gives them the capacity to produce growth miracles but seems to get those miracles to happen very frequently? There can be no simple answer, indeed, any proposed answer is bound to leave out key features, ranging from political changes, evolution of religious beliefs and even historical accident. However, here it will be argued that two features of our economy have played a crucial role. The first such feature is free competition, that is, competition not handicapped by tight government regulations or closely enforced customary rules, like those of the medieval guilds, which prevented gloves-off combat among rival firms. The second crucial development is the fact that in today's economy many rival firms use innovation as the main battle weapon with which they protect themselves from competitors and with which they seek to beat those competitors out. The result is like the case of two countries, each of which fears that the other will attack it militarily and therefore feels it necessary always to match the other country's military spending. Similarly, either of two competing firms will feel it to be foolhardy to let its competitor outspend it on the development and acquisition of battle weapons. Each is driven to feel that at least matching effort and spending on the innovation process is a matter of life and death. Naturally, in an economy in which this is so, a constant stream of innovations can be expected to appear, because firms do not dare to relax their innovation activities.

## 11.6 Innovation versus price as the competitor's prime weapon

There are substantial sectors of the economy in which it seems quite clear that the weapon of choice for competitive battle is innovation. Price does, of course, matter, but it is improvement in processes and products that capture the attention of management. In product lines as diverse as computers and computer software, automobiles, cameras and productive equipment, models are constantly improved, and the improvements are instantly and widely advertised. The firm that can come up with a model better than those of its rivals will have an advantage that cannot be matched by the latter as quickly and easily as a price cut, and certainly the advantage of a dramatic new product is likely to be more substantial. In all of the economy's 'high-tech' industries this appears to be true, and the relationship probably plays a role in many others.

Here, one must not exaggerate. There is strong evidence that most innovation is contributed by a very few industries in a very small number of countries. It is reported that about 80 percent of industrial outlays on new product research and development comes from the chemical and machinery and equipment sectors of manufacturing. Yet, even in many of the industries where product and process development may not be the leading instruments of competition, management cannot afford to neglect them or to leave them to chance. For if one firm fails to do enough in this arena – if it delays in adopting the latest technology, even technology created by others, it becomes far easier for rivals to get ahead of it, very possibly with disastrous consequences for its sales.

As a result, at least in the high-tech sectors of the economy firms do not dare to leave their innovation to chance. Rather, the pressures of the competitive market force them to systematize the innovation process and to seek so far as possible to remove risk from the undertaking. Nowadays this drives business firms systematically and routinely to determine the amounts they will invest in the research and development (R and D) process, systematically decide on the ways in which they will promote and price their innovation and even systematically determine what it is that the company's laboratories should invent.

This kind of business firm innovation activity is far easier to analyze using the tools of the microeconomics of the firm and the industry than it is to analyze what can be described as the 'Eureka!' process, in which a lone inventor working in a basement or garage happens to come up with a brilliant invention. Business innovation is easier to analyze because the decision process related to this activity has become routine, carried out in a manner with much in common with other decisions of the firm, such as how much of one of its output commodities to produce, how much to spend on advertising, and so forth.

## 11.7 How much will the profit maximizing firm spend on innovation?

The basic story I am telling is that competition forces many firms in the economy to keep up their expenditure on research and development, the activity that creates the company's inventions and prepares them for market. While some of the money spent on R and D will be a failure, other such spending will be spectacularly successful, and most R and D outlays will yield modest advances. Taken as a whole, then, one can expect that the more firms spend on R and D, the greater will be the number of innovations that contribute to the economy's GDP. The key questions, then, are how much can we expect firms to spend on R and D, and how will competition affect that amount?

The obvious answer that the firm will act to maximize its profit, spending to the point at which expected marginal revenue equals expected marginal cost, simultaneously tells us everything and nothing about the R and D decision. It tells us nothing about the shapes of the relevant functions or, in particular, how the relationships are affected by competition which, I have suggested, plays an important role in the R and D decision. Nor has the discussion told us what insights we get from the analysis about business behavior and its contribution to economic growth.

## 11.8 The profits of innovation

Many discussions of innovation start off with the assumption that innovators expect to earn very high profits. And this is obviously true of those innovators who create unusually successful innovations. We have all heard of innovators like Thomas Edison, Alexander Graham Bell and, more recently, Bill Gates and others in the computer industry, who have acquired great riches from their ability to invent, bring the innovations to market and sell their products. But for every successful innovator there are many others who have plowed the family savings into their new gadgets, and lost all they have spent. It is possible that on average, inventors have earned zero profits or even less.

Now, entry into innovation is not perfectly easy, but it is much easier than entry into many other economic activities. That means we cannot be certain that economic profits to invention will tend exactly toward the zero level, but we can expect them to be very low on average. In other words, while invention activity will sometimes pay off enormously well, there will also be big failures, so that the average comes out close to zero. This is particularly likely to be true of a large firm that has a big R and D division that simultaneously works on many possible innovations. The law of large numbers makes it very likely that some of these efforts will fail and that some will succeed. Thus, zero economic profit from innovation activity of the firm, that is, profit no greater than what is currently usual in the competitive industries of the economy, is to be expected to be a frequent occurrence in industries with a great deal of innovative activity.

Does this conclusion fit in with the facts? We have no systematic study for all inventive activities. But the high-tech industries do provide a useful case study, and this is particularly true of computers, because it is an industry in which many fortunes have been made and have received much publicity. Here is one report:

> The computer industry hasn't made a dime . . . Intel and Microsoft make money, but look at all the people who were losing money all the world over. It is doubtful the industry has yet broken even,' said Peter Drucker in a recent interview . . . but is it true? Paul Gompers of the Harvard Business School and Alon Brav of the University of Chicago . . . looked at companies that went public from 1975 to 1992, most of which were high-tech firms, and found their rate of return to be about average [i.e., zero economic profit], once they adjusted for risk and company size ('The Rewards of Investing in High Tech,' Federal Reserve Bank of Boston, *Regional Review*, Vol.6, Fall 1996, p. 14).

## 11.9 Risk reduction through technology sharing

We have just shown that low average profits from innovation are in good part attributable to the riskiness of the activity – the high likelihood that effort devoted to an apparently promising innovation will not pay off, and the investment will go down the drain. Firms can be expected to try to minimize the risks, and they do. For example, management may try to avoid providing money to their research laboratories to finance the development of those inventions that are judged to be impractical. There are, however, many examples where management's foresight has not proved to be brilliant – as, for instance, when the Western Union Telegraph Company turned down the newly invented telephone.

Firms also often try to protect themselves by going into partnerships, research joint ventures, that enable them to share the risks, with each of, say, five partner-firms supplying only one fifth of the funds needed to bring an innovation to completion. What is less well known is that many firms try to reduce their risks by systematic technology trading. It is widely believed by those who have not studied the matter that when a firm succeeds in producing a promising new invention it will generally try to keep its competitors from getting hold of it, in order to retain a competitive advantage over its rivals. But that is often not true. Fearing that its own laboratories may conceivably come up with only failures in the year 2003, while its competitor may possibly have better luck that year, a firm will often choose to sign an agreement with that competitor in which each shares with the other all of its successful future innovations, say, for the next five years. This helps to cut down other risks for the two technology-sharing firms. In photography, for example, one camera manufacturer may introduce an improved automatic focus device, another an automatic light adjustment, and a third may invent a way to make the camera lighter and more compact. Each of these three firms has the choice of keeping its invention to itself. But if two of them get together and agree to produce cameras combining the features each of them has contributed, they will be able to market a product that is clearly superior to what each could have produced alone. They are then likely to be in a far better position to meet the competition of the third camera manufacturer.

There are many firms and industries that engage in this practice, industries ranging from steel production to computer manufacturing. The exchanges may be entirely informal, or they may be based on detailed contracts, even requiring each firm to train technical experts from the other in the use of the new technology, and even specifying whether the company requesting such training will pay the travel and living expenses of these experts.

Indeed, the activity of business firms in providing their technology to others, for a profit, has become so commonplace that the Massachusetts Institute of Technology runs a seminar for business firms, teaching how they can be more effective in the technology provision business.[4]

## 11.10 A kinked revenue curve model of spending on innovation

The discussion so far leaves a basic question unanswered. If innovation takes much effort and money, if it is so risky, and if the economic profits to be expected from innovation activity are near zero, why do firms do it? Why does not every firm refuse to join this unattractive game? The answer, at least in part, is that the competitive market mechanism gives them no choice. If they do not keep up with their competitors in terms of attractiveness of their products and efficiency improvements that permit them to keep their costs low, they will lose out to their rivals, and end up losing market share and losing money. Zero economic profits, that is, profits that yield normal competitive returns to investors, surely are better than negative profits.

The result is like an arms race between two countries, each of which fears invasion by the other. Each is driven to keep up with the other's military expenditure. Raising its armaments expenditure will probably get it nowhere, because it can expect the other nation to match any such increase, raising expenditure without improving the nation's military security. But, at the same time, neither nation will dare to cut its arms spending unilaterally, since that will simply invite invasion by the other.

This story can be made more explicit with the help of a microeconomic model very similar to the well-known kinked demand curve model of monopoly pricing that has been proposed to explain why in oligopoly markets prices tend to be 'sticky' because no firm dares either to raise or to cut its price. It will be recalled that the underlying mechanism is an asymmetry in the firm's expectations about the behavior of its competitors. The firm fears either to lower its price below that of its rivals, or to raise it above theirs. It is afraid that if it lowers its product price its rivals will match the price cut, so that our firm will end up with hardly any more customers. On the other hand, if it increases its price it fears that the others will not follow, so that it will be left all by itself as the seller of an overpriced product and will lose all its customers. The result is that normally such a firm will set its price at the industry level, no more and no less, and leave it there unless there is a major change in cost or demand or some other extreme change.

The innovation story is similar. Consider an industry with, say, five firms of roughly equal size, and that the firm with whose decision we will be concerned, Company X, sees that each of the other firms spends about $20 million a year on R and D. X will not dare to spend much less than $20 million on R and D itself, because if it does so its next year's product model will probably not be nearly as good as those of some or all of its rivals. On the other hand, it sees little point in raising the ante, say, to $30 million, because it knows the others will feel themselves forced to raise their R and D budgets correspondingly.

The story can described graphically in a figure that shows the MC curve and a kinked MR curve as functions of the firm's total R and D expenditure. The shape of the MC curve does not matter in this story. The MR curve, however, has a vertical break at the, say, $20 million level of R and D investment that currently is the norm in the industry. The story behind this curve is simple. Consider any lower level of R and D spending by Company $X$, say, $5 million per year. Then its products will grow exceedingly inferior to those of its rivals, but it may sell a few because they meet the special needs of a few customers at highly discounted prices. More R and D spending will bring in a small additional revenue and still more increases in R and D will add more to revenue. The peak will occur at the point where Company $X$ meets the industry standard and its product becomes really competitive with those of its rivals. However, further spending is not very revenue producing because if Company $X$ inflates its R and D budget further its rivals will feel threatened, and match the increase. Thus, further increases in R and D spending by Company $X$ yield a very low MR, as shown by a low segment of the MR curve to the right of the $20 million spending level. The MC and MR curves meet at a point in the vertical part of the MR curve, so it will pay Company $X$ to follow industry practice, investing $20 million a year in R and D, at which MC = MR. It may go on doing so, year after year.

But that is not the end of the story. All five firms in the industry will continue to invest the same amount, until some year one of them has a research breakthrough and comes up with a wonderful new product (as happens in most high-tech industries from time to time). Then, for that firm it will pay to expand its investment in the breakthrough product, because that will pay off even if the other firms in the industry match the increase. The MR curve for that breakthrough firm will rise, its MC = MR point will move to the right, say to $25 million. Other companies in the industry will feel forced to follow, and now the industry norm will no longer be a $20 million investment per year, but will instead be $25 million per firm.

The story, then, is that competition forces firms in the industry to keep up with one another in their R and D investment. But once they have caught up, the investment level remains fairly level until, from time to time, something induces one firm to break ranks and increase its spending, with all the other firms following behind. Such an arrangement is described as a 'ratchet' (in analogy with the mechanical device that prevents a spring that is being wound up from suddenly unwinding). It is an arrangement that holds matters steady, permits them under certain circumstances to move forward, but generally does not allow them to retreat. R and D spending can then be expected to expand from time to time, but once the new level is reached, the ratchet – the competitive market – prevents a retreat to the previous lower level.

This, in my view (based on considerable but unsystematic observation), is a critical part of the mechanism that accounts for the extraordinary growth record of free-enterprise economies and differentiates them from all other known economic arrangements. It is the competitive pressure that forces firms to run as fast as they can in the innovation race just in order to keep up with the others.

## 11.11 Three growth-creating properties of innovation

We have just seen reasons to expect that the market mechanism will force firms to devote at least a steady stream of resources to innovative activities, notably to R and D. With luck, such a level R and D effort will yield a fairly level flow of innovations. But a level flow of innovation does not mean that GDP will remain level. Rather, a *level flow of innovation* can be expected to result in *steady growth of the economy's output*. Here we must take note of three critical features of innovation that can, so to speak, magnify the contribution of technical change to the economy's GDP. These three features are (i) the cumulative character of many innovations, meaning that many innovations do not merely replace older technology and make that technology obsolete; rather, they *add* to what was previously available, thus constituting a net increase in the economy's inventory of technical knowledge. (ii) In addition, an innovation, once created, need not contribute only to the output of the firm that made the breakthrough. At little or no additional cost it can also add to the outputs of other enterprises. This is, of course, the public good property of technical knowledge. (iii) Finally, there is what can be called an accelerator feature of innovation – a level stream of innovation usually means that output will not be level, but growing. It is like stepping on the accelerator, where a steady unchanging pressure makes an automobile move, that is, change position, more rapidly. It is only the last of these attributes that may be unfamiliar to readers, and so a few further words on the subject may be appropriate.

## 11.12 The accelerator property of innovation

It should be clear that each successful innovation adds to the nation's GDP by permitting more products to be created with a given quantity of resources (a 'process innovation') or by making new and more valuable products available ('product innovation'). Thus, an economy whose R and D produces a steady and *unincreasing* output of one innovation per month will obtain a GDP that is *higher* each month than it was in the previous month. That is to say, the economy's output will *grow without letup* even though the flow of innovations that fuels that output growth remains steady at one invention per month. This acceleration relationship applies to innovation generally, so that if the competitive market mechanism were only to lead firms to supply a level quantity of resources to R and D, we would expect continued growth of GDP to result. Of course, as noted, the ratchet principle tends to increase the expenditure of resources on innovation, and not just to leave them level. The acceleration principle tells us about the effects of this too. It tells us that if the level of R and D spending were to increase just once, for example, and stay at that new higher level forever, then the growth rate of GDP would also move to a higher rate, and GDP would increase at this new faster rate forever.

All of this reinforces the role of the competitive market mechanism and its stimulation of innovation as a contributor to the extraordinary growth that characterizes the world's free-enterprise economies.

## 11.13 Free-enterprise growth: routine versus independent endogenous innovation

The analysis of this chapter has focused on the innovation activities of large business firms, their routinized character and the influence of the competitive market forces on the magnitude of such routine innovative activities of the firm. The size of such activity is substantial, and its funding can be estimated to amount to some 70 percent of R and D expenditure in the United States.

Still, a good deal of important innovative activity takes place outside the large corporation. Indeed, there is evidence that a very considerable portion of the innovative breakthroughs of the twentieth century are to be attributed to such independent innovation activities. Nothing in my discussion is meant to minimize their importance. They have not been emphasized here only because they are not so directly influenced by the market mechanism, and therefore probably play less of a role than routine innovation activity in explaining the growth record of the free market – which is my main focus here. Of course, even independent innovation is subject to endogenous influences. For one thing, the success of one innovator is likely to encourage the activity of another and to make it easier for the latter to obtain funding. Nevertheless, independent invention activities have occurred in abundance in some non-capitalist economies, as we know. They are part of my story, but important though they may be for society, they do not seem to be at the heart of the explanation of the growth record of free-market economies.

## 11.14 On the efficiency of innovation activity in free markets

As already noted, standard microtheory suggests that the economy's innovative activity is apt to be extremely far from optimal. The prime reason for this is the spillovers of innovation – their presumably beneficial externalities, which on standard grounds lead us to expect that investment in innovation will be less than optimal. The second apparent shortcoming of the process is the proprietary behavior of innovating firms that leads them to resist sharing their knowledge with others, thereby condemning rivals (and others) to use obsolete methods and to provide obsolete products. Here, to keep my discussion from growing too long, I will only hint at my analysis of these issues. With regard to the second issue – unwillingness to share technology – I have already indicated that market behavior is often very different. Rather than struggling to prevent others from obtaining their proprietary information, firms often trade it for access to the proprietary information of other firms, including competitors. They are often forced to do this by market pressures because firms that share information are likely to end up with better products, benefitting from the combined innovations of all the participants, than the products of firms that rely on just their own innovative resources. The amount of sharing that results may or may not approximate optimality, but it is surely better than a world where secrecy, patents and other influences systematically prevent much of economic activity from use of the latest technology.

Many economists also, apparently, believe that the market does not induce private firms to invest the socially optimal amount in innovation. They believe that many innovations whose benefits would exceed their costs are never carried out by industry, because the firm that spent the money to produce the innovation would get only part of the benefit, a part insufficient to cover the cost. This is why a good deal of innovation and research activities are financed by governments and carried out by research institutions such as universities. This is particularly true of *basic research*. However, it does not follow that the market's performance on quantity of resources devoted to innovation is as far below the optimum as might have been believed. The reason is that there is a tradeoff between increased flow of invention and the distribution of benefits because of which zero externalities cannot be expected to be optimal. Moreover, there is no one level of expenditure that is unambiguously optimal. Instead, there is a range of values of what I call the *spillover ratio*, that is, the share of the benefits of innovation that goes to persons other than the investors, such that *all values of the ratio within this range are Pareto-optimal*. Consequently, there is no way in which economic analysis alone can choose among them. Rather, value judgments must be employed in making that selection. Indeed, it can even be suspected that the high spillover ratio found in reality falls within the range of Pareto optimality.[5] The reason for these results is that, in contrast with most of the literature (except that on optimal duration of patents), here spillovers are considered to be capable of offering social benefits, and do not always simply impede or prevent the attainment of optimality. The point is that there is an inevitable tradeoff between the number of innovations actually produced and the standard of living of the majority of the population. In this scenario, as overall GDP is raised, any increase in workers' standards of living constitutes a rise in the spillovers from innovation that depresses the flow of further innovation. Thus, the more the general public benefits from such growth in GDP, the slower that growth must be.

This is more than just an embellishment of the old story of the tradeoff between output and distributive equality. The mechanism under discussion here is very different, and does not involve the disincentive to work that results from a reduction of the marginal return to worker effort. Rather, we are concerned here with the heart of the capitalist growth process: the payoff to innovation and the speed with which new technology and new products become available.

Our scenario is by far the more dramatic. Romer notes in passing that, if the innovator were totally immune to the disincentives of spillovers, then none of the benefits would have gone to others. But, if that were so, then real wages would hardly have risen from their levels before the Industrial Revolution![6] It is almost impossible to imagine how great a difference that would have entailed. If we assume the most extreme case – that the spillovers from innovation are reduced to (anywhere near) zero – the living standards of the vast majority of the citizens of today's rich countries would have stalled at pre-Industrial Revolution levels. One can hardly accept the notion that it would be socially preferable to achieve a total GDP that is far higher than today's through enhanced incentives for innovation, while the bulk of the population is condemned to near-medieval living standards, but that is where such a premise leads us. Even the fortunate few innovators who might amass unimaginable wealth in such a zero-spillover world would undoubtedly prefer somewhat better conditions for their impoverished compatriots, and would themselves probably be better off not only in terms of the social environment, including reduced violence and disease. In addition, innovators would probably have higher absolute incomes because more can be produced by a labor force that is better fed, healthier and better educated. But any such gain to labor is unavoidably a rise in the *percent* of the return to innovation that goes to others than the innovator – it is necessarily a rise in the share of spillovers.

Once again, my conclusion is not that the market's performance in response to the externalities of innovation is optimal. I suggest only that it is far better than standard theory might lead us to expect, and comes closer to consistency with the observed and unprecedented growth record of the free-enterprise economies.

## Acknowledgement

I am grateful to the Russell Sage Foundation and the C. V. Starr Center at New York University for their support of this work.

## Notes

1 According to the late British economist, T.S. Ashton, in his classic book, *The Industrial Revolution, 1760–1830*, London: Oxford University Press, 1948, that phrase is how one schoolboy (quite appropriately) described the Industrial Revolution.

2 See R.E. Lucas, Jr (1988) 'On the Mechanics of Economic Development' *Journal of Monetary Economics*, 22, 3–42

3 For a discussion of possible reasons for failure of the economies of Ancient Rome and Medieval China to achieve outstanding growth records, see my *Entrepreneurship, Management and the Structure of Payoffs*, Cambridge, Mass.: MIT Press, 1993, Chapter 2.

4 For further materials on technology sharing in practice, with a number of concrete examples, see Chapter 10 of my recent book (op.cit., 1993).

5 My own value judgment is summed up in George Bernard Shaw's dictum that there is no crime greater than poverty. Consequently, I am inclined to prefer a fairly high spillover ratio, perhaps not far from its current value.

6 'This pattern of industrialization without wage gains is what it would take to ensure that the industrialist captures all of the benefits he creates when he introduces machinery. . . . [this] cannot be a historically accurate description of the process of development in industrial countries, for if it were, unskilled labor would still earn what it earned prior to the industrial revolution' (Paul Romer, 'New Goods, Old Theory and the Welfare Costs of Trade Restrictions,' *Journal of Development Economics*, Vol. 43, 1994, p. 29).

# Part IV

# Challenges for quantitative methodologies

# 12 Business cycle research

## Methods and problems

*Edward C. Prescott*

### 12.1 Methods in business cycle theory

Basketball and soccer are different games. The objective is the same, namely to put the ball in the net. In both games, a ball is dribbled and passed. However, what is good dribbling and passing practice when playing one game is bad practice when playing the other. A soccer player who picks up the ball and bounces it as he runs down the field is not following good practice. This activity, however, is good practice when playing basketball.

In scientific inference there are two fundamentally different and complimentary games. The objective of both is the same, namely to draw scientific inference. One game is drawing deductive inference. The other game is drawing inductive inference. Both require the selection of a model. In this way they are similar. However, the rules for selecting the model are fundamentally different in these games.

This distinction between inductive and deductive scientific inference is important in business cycle research. Business cycle theory uses deductive or quantitative theoretic inference. Often practitioners do it poorly. They do it poorly because they use practices that are good for inductive inference but bad for deductive inference. Their behavior is analogous to the soccer player who picks up the ball and bounces it as he runs down the field.

I first clarify the differences between these forms of scientific inference. I do this by reviewing the role played by each in the development of the natural sciences. With inductive or empirical inference *the product is the model or the law*. In physics an example of inductive inference is Kepler's laws of planetary motion. These laws are:

- Law 1 – Planets follow an elliptical orbit around the sun,
- Law 2 – The sun is at one of the foci of each ellipse,
- Law 3 – An equal area of a planet's ellipse is swept out in a given time interval by the planet.

Another example of successful inductive inferences in the natural sciences is Galileo's discovery of the law of motion or model of a ball rolling down an incline plane. This law is:

$$D = 1/2g\ \sin(\theta)t^2,$$

where $D$ is distance, $t$ is time, $\theta$ is the angle of the inclined plane, and g is a constant. An important feature of the formula is that the weight of the ball does not appear. I do not use the term mass as that concept did not exist when Galileo discovered his law. The concept of mass had meaning only subsequent to the development of Newton's theory of mechanics.

Inductive or empirical inference was productive in the natural sciences. Good inductive practices involved model estimation and testing. This raises the question of why inductive or empirical inference proved sterile in business cycle research. This sterility was not due to the incompetence of the researchers who pursued the inductive approach. The group who pursued this research program included a disproportionate number of the best minds in economics. The reason these inductive attempts failed, I think, is that the existence of policy invariant laws governing the evolution of an economic system is inconsistent with dynamic economic theory. This point is made forcefully in Lucas' famous critique of econometric policy evaluation.

With deductive inference *a model is a tool or measurement instrument* used to deduce the implication of theory. This statement requires a discussion of what theory is. The definition that I will use is an implicit set of instructions for constructing a model economy for the purpose of answering a question. Two examples of theory drawn from physics are: (i) Newtonian Mechanics – force equals mass times acceleration – and (ii) Newton's Universal Law of Gravitation – the gravitational force operating on two bodies is proportional to the product of the masses and inversely proportional to the square of the distance. These theories provide a theoretical foundation for Kepler's laws of planetary motion and Galileo's law of a ball rolling down an inclined plane. Newtonian mechanics has proven useful. This theory is used to construct models for all kinds of purposes in the engineering sciences. It is used to predict the path of rocket ships and to control their paths. It is used to design machinery in factories.

In economics there is Walrasian general equilibrium theory, which Schumpeter (1954) judged to be 'the only work by an economist that will stand comparison with the achievements of theoretical physics'. I do not agree with the 'only' part of Schumpeter's statement, as I would add Arrow–Debreu general equilibrium theory to this list, and possibly game theory and the closely related mechanism design theory.

General equilibrium theory is theory in the language sense but not in the sense that I am using the word theory. The reason is that without restrictions on preferences and technology general equilibrium theory is virtually vacuous. Consequently it is not a set of instructions for constructing an instrument to measure something or predict the consequences of some policy. Growth theory with measures of the elasticity of substitutions and transformations and share parameters is theory in the sense I am using it here. Growth theory provides instructions for constructing a model economy to address some question of interest. The quantitative answer to the question is *deduced* for the model economy. In business cycle studies, growth theory is the theory used. Indeed, business cycle research is largely drawing inference from growth theory for business cycle fluctuations. Growth theory is also heavily used to construct models to estimate the welfare and other quantitative effects of tax policies and social security systems.

In this review I will explain why using estimation theory to select a model used in drawing deductive inference is bad practice. Unfortunately, or maybe fortunately for economists, there is no set of mechanical rules for selecting a good model for deducing some inference of a theory. I say it may be fortunate because, if it were mechanical, computers could replace economists. I now illustrate this point by reviewing the development of modern business cycle theory.

### *12.1.1 History and overview of business cycle theory*

Burns and Mitchell (1946) developed a statistical definition of a business cycle. This definition did not prove useful. Kuznets (1946a, 1946b) and his students were more successful. They systematically reported economic events using a system of national income and product accounts. They also reported aggregate factor inputs of capital and labor. A set of growth facts emerged from this reporting. These facts guided researchers in the development of growth theory. Some of these facts are as follows. Labor share of product, using Kravis' (1959) economy-wide assumption, is more or less constant over time even though the real wage increased dramatically relative to the rental price of capital. Subsequently Gollin (forthcoming) found that this regularity held across countries with deviations being unrelated to the level of development. Another fact is that the investment share of product is more or less constant. This led to the Solow growth model with a Cobb–Douglas aggregate production function. The Solow model with its aggregate production function and factors being paid their marginal product is a theory of the income side of the national income and product accounts.

The Solow growth model, with its exogenously determined savings rate, led the economic theorists Cass (1965), Koopmans (1965) and Diamond (1965) to develop a theory of the allocation of product between consumption and investment. Brock and Mirman (1972) extended this theory to stochastic environments.

Lucas (1977) defined business cycles to be recurrent fluctuations of output and employment about trend. He wrote that the key business facts were the co-movements of the economic time series. Hodrick and I (1980) developed a statistical definition of the business cycle component of an economic time series. The regularities that appeared are all tied to the variables in growth theory. They are: (i) consumption is strongly procyclical and fluctuates about a third as much as output in percentage terms; (ii) investment is strongly procyclical and fluctuates about three times as much as output; (iii) two-thirds of output fluctuations are accounted for by variations in the labor input, one-third by variations in TFP and essentially zero by variations in the capital input; (iv) the only important lead–lag relation is that the capital stock lags the cycle with the lag being greater the more durable the capital good; (v) the deviations of output from trend, that is the business cycle component, displays a moderately high degree of persistence; (vi) the real wage is procyclical but is roughly orthogonal to the labor input.

These facts were bothersome for theorists. Why should leisure be low when consumption is high? After all, consumption and leisure are normal goods and leisure is not high when the real wage is high. Another question is why labor productivity is high when labor input is high. This violates the law of diminishing returns.

Many have argued that Hodrick and my facts are not interesting because we did not correctly measure the business cycle. This criticism is spurious. An operational definition can be neither right nor wrong and our definition is an operational definition. Economics is not the only science where operational definitions have proven useful. In the natural sciences prior to the development of the theory of an ideal gas the definition of temperature was an operational one being neither right nor wrong. An environment was 50 degrees Celsius if the thermometer registers half way between what it registers when it is immersed in ice water and what it registers when immersed in boiling water.

In retrospect, Hodrick and my representation of time series as the sum of two components, one we called the growth or trend component and the other the business cycle or deviation component, turned out to be a useful one. Our representation revealed some behavior that was in apparent contradiction with theory and fostered the development of some good theory.

I emphasize that these facts were in *apparent* contradiction with theory. Until the dynamic applied general equilibrium tools were developed to derive the implication of growth theory for business cycle fluctuations, economists had to rely on their intuition derived from price theory. This price theory intuition proved to be wrong.

Exploiting Arrow–Debreu language, recursive methods, and computational methods, Kydland and I (1982) derived the implications of growth theory for business cycle fluctuations. To our surprise we found that, if total factor productivity (TFP) shocks are persistent and of the right magnitude, business cycle fluctuations are what growth theory predicts. Subsequently I (1986) found that these TFP shocks are highly persistent and of a magnitude that implies that they are the major contributor to business cycle fluctuations. This success of growth theory led me to have greater confidence in public finance findings that use growth theory to evaluate tax policies.

Kydland and Prescott (1982) examine the consequence of people valuing leisure more if they have worked more in the past. The introduction of this feature into growth theory preserved the consistency of the theory with the growth facts. Its introduction increases the inter-temporal elasticity of substitution for leisure and results in the prediction of the model being more in conformity with aggregate observations. However, this feature should not be part of the measurement instrument used to answer the question of how volatile the US economy would have been if TFP shocks were the only shocks until other evidence is provided that it indeed quantitatively describes people's preferences. Such evidence never materialized. This leads to the following two principles for selecting the model used in business cycle research.

**Principle 1:** *When modifying the standard model of growth theory to address a business cycle question, the modification should continue to display the growth facts.*

**Principle 2:** *The model economy being used to measure something should not have a feature which is not supported by other evidence even if its introduction results in the model economy better mimicking reality.*

When modifying the standard growth model to address a business cycle question, Hansen (1985) introduced another feature, a labor indivisibility, and permitted people to enter into mutually beneficial insurance contracts as in Arrow–Debreu theory. This increases the magnitude of the response to TFP shocks. There is empirical evidence that justifies incorporating this feature. First, there is unemployment insurance and people have a close substitute for insurance, namely the holding of liquid assets. Second, and most important, most of the cyclical variation in the labor input is the result of variation in the number of people that work in a given week and not in the average workweek length. Like Kydland and my non-time-separable utility function, the introduction of this feature results in leisure having a high intertemporal elasticity of substitution.

Hansen finds that for his model economy, cyclical fluctuations induced by technology shocks are as large as the observed fluctuations. Given the strong empirical support for labor indivisibility, this mechanism should be part of the model or measurement instrument used to answer the question of the importance of technology shocks for business cycle fluctuations. A problem, however, is that not all fluctuations in hours are the result of the variations in the number of people working. An important fraction is the result of variation in the workweek length as in Kydland and my model economy. This number is somewhere between the Hansen and the original Kydland and Prescott estimates.

What was needed was better theory. The reason is the following. Why is the workweek fixed? If it is permitted to vary and aggregate hours is the labor input to production, then all work and all variation is in the workweek length. There was a major inconsistency between observation and theory. Better theory was developed that reduced this inconsistency. I begin with the aggregate production function and its inadequacy for the purpose of understanding the determination of workweek length.

### *12.1.2 The aggregate production function*

The aggregate production function with labor and capital as its inputs, which has proven so useful in public finance, fails to capture the intensity with which people and machines work. These intensities vary over the cycle. Before discussing how to extend the theory of the production function utilizing results of Mas-Colell (1975), Hart (1979), and Jones (1984) on differentiated commodities, the aggregation theory underlying the aggregate production function will be reviewed.

The technologies underlying the aggregate production function:

**(i)** There are $n$ factor inputs and a composite output.

**(ii)** The vector of inputs is $x \in \Re^n_+$ and the output good is $y$.

**(iii)** A plant technology is indexed by $x \in T$ with $f(x)$ being plant type $x$ output.

**(iv)** $X \in \Re^n_+$ is the vector of aggregate inputs and $Y$ aggregate output.

**Definition:** An *aggregate production function* $F(X)$ is the maximum output that can be produced given the input vector $X$.

**Assumption 1:** Any number of technologies of type $x \in T$ can be operated.
**Assumption 2:** For all $x \in T, x$ is infinitesimal relative to $X$.
**Assumption 3:** $T \subset \Re^n_{++}$ and $T$ is compact.
**Assumption 4:** $f : T \rightarrow \Re$ is continuous.

The aggregate production function is the solution to the following program, where $M_+(T)$ is the set of measures on the Borel sigma algebra of $T$,

$$F(X) = \max_{z \in M_+(T)} \left\{ \int f(x) z(\mathrm{d}x) \right\}$$

subject to

$$\int_T x_i z(\mathrm{d}x) \leq X_i \quad i = 1, 2, \ldots, n.$$

Given the assumptions, the constraint set is compact and non-empty and the objective function is continuous in the weak star topology. Therefore the program has a solution. Two well-known results are the following.

**Proposition:** $F(X)$ is weakly increasing, continuous, weakly concave, and homogenous of degree one.
**Proposition:** If there is free entry, profit maximization results in output being maximized.

### *12.1.3 An example*

The Cobb–Douglas production function has come to dominate in aggregate applied general equilibrium analysis. The reason is that both over time and across countries, labor's share of product is surprisingly constant at about 70 per cent.[1] The Cobb–Douglas production function, with its unit elasticity of substitution, is about the only aggregate production function with the property that factor income shares are the same even though relative factor prices are very different.

An example of an underlying set of plant technologies for the Cobb–Douglas production function is the following one. Suppose that the factor inputs are capital $k$ and labor $n$ and that the plant technologies are $g(n)k^{\theta}$ . In addition, the function $g$ is such that the function $g(n)n^{\theta-1}$ has a maximum. This maximum is denoted by $A$ and a maximizing n by $n^*$.

**Proposition:** For this example, the aggregate production function is $F(K,N) = AK^{\theta}N^{1-\theta}$.

*Outline of proof*: Given there are two constraints and an optimum exists, there is an optimum with at most two types of plants operated. Consider one such optimum. Let $(K_i,N_i)$ be the aggregate factor inputs used to operate type $i$ plants for $i = 1,2$. Allocate $(K_i,N_i)$ equally to m plants of type $(K_i/m, N_i/m)$. The $m$ for which $N_i = mn^*$ is an optimum. Thus, all operated plants for this optimum have $n = n^*$. Output maximization requires marginal products of capital be equated across operated plants. Thus, operating $N/n^*$ plants each with $n^*$ workers and $K/(N/n^*)$ units of capital is optimal. Using this result,

$$F(K,N) = \frac{N}{n^*}g(n^*)\left(\frac{K}{Nn^*}\right) = AN^{(1-\theta)}K^{\theta}$$

### 12.1.4 Workweek of capital

The workweek of capital and labor varies cyclically. The aggregate production function does not capture this. The following technologically does:

$$y \leq Ahk^{\theta},$$

where $0 < \theta < 1$ and where $y$ is the output produced by an individual, $k$ is the capital that that individual uses and $h$ is the length of that individual's workweek. Element $h$ belongs to the set $H \subset (0,1]$.

**Question:** What is the aggregate production function for this technology?

A workweek of different length is a different factor input. Following the procedures above, there is an aggregate production function. The inputs are capital $K$ and the measure $N \in M(H)$. $N(B)$ is the measure of people working a workweek belonging to Borel measurable set $B \subseteq H$. An aggregate production function $F(K,N)$ exists.

To simplify the exposition I deal with the case that there are only a finite number of possible workweek lengths. Then $N_h$ is the measure of workweeks of length $h \in H$ and $N$ is a finite dimensional vector. For each workweek length the aggregate production function is

$$F_h(K_h,N_h) = hAK_h^{\theta}N_h^{1-\theta}$$

The aggregate production function $F(K,N)$ is obtained by equating marginal products of capital across these aggregated technologies and summing over $h$.

Finn Kydland and I (1991) introduced this feature into the growth model. In our model economy households' preferences are ordered by the expected value of

$$\sum_{t=0}^{\infty} \beta^t \frac{\left[c_t^{1-\Psi}(1-h_t)^{\Psi}\right]^{1-\sigma}}{1-\sigma}.$$

The technology parameter $\{A_t\}$ is governed by a first order auto-regressive process with high persistence. Capital depreciates exponentially. The parameters are selected so that the model economy displays the growth facts including the rough constancy of the fraction of time allocated to the market, including commuting time.

Kydland and I were surprised that cyclically only the fraction of the population working varied and not hours per employed worker. Hornstein and I (1993) find that this is precisely what theory predicts. The fact that hours per worker do vary, however, is not bothersome for theory. Kydland and I (1991) find tiny costs of moving between the market and the household sector results in $h$ varying cyclically, as it does.

Introducing this option to vary the length of the workweek along with some moving costs results in observations being in better conformity with theory. The costs are selected so that the relative variability of employment and workweek length match observations. The resulting model, or measuring instrument, is a better one than either Kydland and my original model or the Hansen model. Using this model Kydland and I estimated that the US post-war economy would have been 70 per cent as volatile if total factor productivity shocks were the only disturbance. Here the volatility measure is the variance of the business cycle component.

With this 70 per cent number, an implication of theory is that labor productivity should be orthogonal to the labor input, as it is in the data. If the estimate contribution had been near 100 per cent, an implication of theory is that labor productivity and the labor input should be highly correlated, as they are in the model economy. This leads to an important methodological point. Given this 70 per cent estimate, if both the model and the actual economy had high correlation between labor productivity and the labor input, it would have been a basis for rejecting the model as a good instrument for measuring the importance of TFP shocks.

**Principle 3:** *A model that better fits the data may be a worse measurement instrument. Indeed, a model matching the data on certain dimensions can be the basis for rejecting that model economy as being a useful instrument for estimating the quantity of interest.*

**Principle 4:** *A corollary of Principle 3 is that using statistical estimation theory to estimate models used to deduct scientific inference is bad practice. Estimating the magnitude of a measurement instrument, whether it is a thermometer or a model economy makes no sense.*

There are legitimate challenges to this 70 per cent number. Challenging the result because the model used is not realistic is not one. As stated previously all models are abstractions and therefore unrealistic. A legitimate challenge is to introduce some feature into the model economy in a quantitatively reasonable way and show that the answer to the question changes. There have been many such challenges to Kydland and my finding, but the result has been found to be robust. Before proceeding with a review of some of these studies, I will state the fifth and last principle.

**Principle 5:** *A legitimate challenge to a finding is to introduce a feature into the model economy that is serving as the measurement instrument in a quantitatively reasonable way and show the answer to the question changes.*

### *12.1.5 Robustness of results to increasing returns and monopolistic competition*

Hornstein (1993) extends the neoclassical growth model to incorporate monopolistic competition and increasing firm-specific returns to scale. In his model world there is a large and fixed number of firms where each firm has market power for its own product. Increasing returns to scale at the firm level is introduced through a fixed cost to production. This assumption allows for an equilibrium where firms make profits but, on average over time, profits are zero.

To account for the role of productivity shocks as sources of output fluctuations Hornstein focuses on two effects. First, he studies to what extent the Solow residual overestimates actual productivity changes. Second, he demonstrates that, compared with the basic neoclassical growth model, a productivity shock generates a stronger output response and a weaker employment response. The change in output and employment responses can be attributed to the increasing returns to scale that also generates stronger wealth effects and thereby dampens employment fluctuations. Finally, he suggests that the net effect of increasing returns to scale and monopolistic competition is to lower the contribution of the productivity shocks to output fluctuations somewhat, but that this effect is limited when markups and returns to scale are not unreasonably large. Cooley, Hansen and Prescott (1995) find similar results for a model economy with idle capital.

Devereux, Head and Lapham (1996) incorporate technology shocks into a real business cycle model with monopolistic competition and increasing returns to both specialization and scale. They find that market power and increasing returns due to fixed costs have no effect on the responses of aggregate variables to a technology shock when compared to those exhibited by a standard, perfectly competitive real business cycle model. The responses of aggregate variables to technology shocks are actually increased by returns to specialization and reduced by returns to scale in variable factors. They find that returns to specialization and scale also affect the measurement of technology shocks. The variance of the Solow residual understates the variance of the technology shock due to increasing returns to scale. Returns to specialization result in the opposite bias. When both types of increasing returns are present, the authors find the variance of output is *increased* relative to a standard competitive model despite a significant reduction of the variance of technology shocks.

Fagnart, Licandro and Portier (1999) investigate the phenomenon of under utilization of productive equipment and its implications for business cycles. The authors introduce the concept of capacity utilization (as opposed to capital utilization) into a stochastic dynamic general equilibrium model. Monopolistically competitive firms use a 'putty–clay' technology and decide on their productive capacity and technology under idiosyncratic (demand) uncertainty. It is shown that the proportion of firms with excess capacities plays an important role in magnifying and propagating aggregate (technological) shocks. Furthermore, they find that idiosyncratic uncertainty about the exact position of the demand curve faced by each firm explains why some productive capacities may remain idle in the sequel and why individual capacity utilization rates differ across firms. Finally, the variability of capacity utilization allows for a good description of some of the main stylized facts of the business cycle and generates endogenous persistence.

### *12.1.6 Consequence of capital and capacity utilization variation for the estimate*

Greenwood, Hercowitz and Huffman (1988) make the assumption that the intensity with which capital is used, say $x$, is a choice variable. The amount of capital services is the product of $x$ and $K$. The cost of using capital more intensely is that depreciation is greater.

$$K_{t+1} = (1 - \delta(x))K_t + I_t.$$

The function $\delta(x)$ is increasing in $x$.

Introducing this feature results in aggregate observations being in greater conformity with theory. However, there is a problem. The problem is the lack of micro-observations to back up the depreciation assumption. Does capital depreciate more in boom periods? Until other evidence is provided for this depreciation assumption, this feature is best not incorporated into the measurement instrument.

### *12.1.7 Labor hoarding over the business cycle*

Burnside, Eichenbaum, and Rebelo (1993) introduce the possibility of labor hoarding. In their model world, people must commit to how many hours they will work over the next three months. During this three month period only the worker's work intensity can vary. In fact hours of employment are not fixed for such long periods for much of the workforce. There are weekly layoffs and variations in the length of the workweek. Their analysis is important for it nicely puts to rest the labor hoarding story that has confused the profession for so many years. The paper accomplishes this by showing how extreme and implausible assumptions must be for labor hoarding to be a factor in understanding business cycle fluctuations. Labor hoarding is only important when shocks are temporary, not when they are highly persistent, and business cycles are responses to highly persistent shocks.

## 12.2 Problems in business cycle theory

There is no shortage of important open problems in business cycle theory. What is in short supply are problems that are both important and analyzable using existing tools. My view is that whenever new tools are developed, it is a good time to search the set of important open problems for one that can be analyzed using these new tools. With this in mind I focus only on problems for which the needed tools have been recently developed or are being developed.

### *12.2.1 The role of organizations in business cycles*

Fitzgerald (1996, 1998) develops and uses a general equilibrium framework in which the number of hours and the employment levels of heterogeneous workers is endogenously determined. He does this in an environment where production requires coordinating work schedules of different worker types, a characteristic that he refers to as team production. In particular, he assumes that all workers in a production team must work the same hours. Output is produced by a large number of teams, where team composition and the hours a team works can differ across teams.

He has two types of workers, the skilled and the unskilled. Skilled workers lose human capital if they are not employed. He finds that constraints on workweek length increase the welfare of the high income skilled people and reduce the welfare of the low income unskilled workers. Furthermore, he finds there is an increase in the employment rate of the unskilled workers. He also finds that introducing this realistic feature does not change the estimate of the importance of total factor productivity shocks for business cycle fluctuations.

In the model he uses preferences that are additively separable in consumption and leisure, namely

$$E\left\{\sum_{t=0}^{\infty}\beta^t[\log(c_t(s))+\nu(1-h_t(s))]\right\}.$$

A suitable commodity vector is $\{c_t(s),i_t(s),k_t(s),n_{1t}(s),n_{2t}(s)\}_{s\in S,t\in\{0,1,2,\ldots\}}$. Here $s$ is the state, $c$ and $k$ are scalars, and the $n_i$ are signed measures on the Borel sigma algebra of [0,1]. Restrictions on a type $i$ consumption set are that $n_i$ be a probability measure and $n_{j\neq i}$ be the null measure.

The plant production function is

$$hf(k,x_1,x_2)$$

where $h$ is hours the plant is operated, $k$ is the capital employed and the $x_i$ are the number of type $i$ workers. Suppressing the $t$ and $s$ index, a period aggregate production possibility set is

$$
\begin{aligned}
Y = \quad & \{(C,I,K,N_1,N_2,) \geq 0 | \exists \text{ measure } z(\mathrm{d}h \times \mathrm{d}\,k \times \mathrm{d}\,x_1 \times \mathrm{d}\,x_2) \text{ for which} \\
& C + I \leq \textstyle\int f(k,x_1,x_2,)\mathrm{d}z \\
& \textstyle\int k\,\mathrm{d}z \leq K \\
& \int_{\{(h,k,x_1x_2)|x_i \in B\}} \mathrm{d}z \leq N_i(B) \text{ all measurable } B \text{ all } i\}
\end{aligned}
$$

The set is a convex cone so there are constant returns to scale.

### *12.2.2 The role of money in business cycles*

A question that has received a great deal of attention is whether monetary factors are a major contributing factor to business cycle fluctuations. Friedman and Schwartz (1963) argued in a comprehensive empirical study that monetary shocks are the major cause of business cycle fluctuations. They observed that sharp declines in the money stock occurred prior to severe economic downturns.

The apparent inconsistencies of economic fluctuations with economic theory that abstracted from money led to widespread acceptance of the Friedman–Schwarz view even though a theoretical foundation was lacking. Real business cycle theory finds that a major fraction of US postwar business cycle fluctuations is accounted for by persistent shock to total factor productivity (see Kydland and Prescott (1991). If money is not a major contributing factor to business cycle fluctuations, why is money highly correlated with output? Freeman and Kydland (1998) provide a possible answer using a transaction based theory of money. They introduce a costly transaction technology into the standard growth model and find that monetary aggregates and output are positively correlated even though there is no *causal* relationship between money and output.

An issue is how to model money. Currently the dominant view is that money is valued because it facilitates existing trade or permits new trade. Lucas (1980) and Lucas and Stokey (1987) develop cash-credit goods and cash-in-advance models. Ireland (1995), Schreft (1992), Freeman and Kydland (2000) develop transaction based models. Saving (1971) and McCallum and Goodfriend (1987) develop shopping time models. Christiano and Eichenbaum (1995), Fuerst (1992) and Scheinkman and Weiss (1986) develop and explore limited participation models.

A question is how do these transaction-based theories stand up to the facts. At the very low frequencies their performance is impressive (see Lucas (1988)). At the very high frequencies there are no problems because, with Lucas surprises, these theories place weak restrictions on observations. At the intermediate frequencies, these theories with their empirical demand for money relations fail and fail spectacularly. These theories predict much larger variations in the demand for money than observed given the variation in the nominal interest rate.[2]

Whether this failure is a serious one for evaluating the importance of monetary shocks for business cycle fluctuations is an open issue. This question, however, is likely to be central for evaluation and design of monetary stabilization policies, which is the principal reason for developing a theory of money. Perhaps the empiricists are right that adopting a constant money growth rule will result in less fluctuations in output and employment. Perhaps they are wrong. Until there is some strong monetary theory the answer is unknown. Once there is a good theory for evaluating monetary policy, the welfare benefits and costs of alternative policy rules can be deduced.

Researchers using the RBC methodology and a transaction based theory of money have found that money contributes little to business cycle fluctuations. I, however, see these exercises as being far from conclusive in establishing that monetary policy cannot be used to stabilize the economy. The reason is the failure of these theories at the intermediate frequencies. Diaz-Gimenez, Prescott, Fitzgerald and Alvarez (1992) have taken an alternative approach that may not suffer from this deficiency. Key in their model is the government issuance of nominal debt that is held directly or indirectly by households. In the United States in 1986 they report that the outstanding stock of nominal debt is nearly one annual GNP. There is also a large amount of lending by old to young in nominal terms to finance home purchases. Consequently there is a lot of risk associated with unexpected inflation that results in redistribution between young borrower and old lenders. A question is why markets do not develop to eliminate this risk. One answer is that there is no way for the unborn to contract with people alive to avoid the risk associated with this redistribution.

Another possibly important component of a successful theory for evaluating monetary policy rules is intermediation costs. These costs are large. In rich countries the spread between the average borrowing and average lending rates is about 5 per cent. The total product of the financial sector was 9 per cent in the United States in 1986. Still another possibly important component of a successful theory for evaluating monetary stabilization policy is moral hazard associated with allocating idiosyncratic risk. Empirically much of this idiosyncratic risk is not eliminated, which would be the case if financial markets were frictionless and there was a complete set of state contingent commodities.

### *12.2.3 The role of policy in determining labor-leisure time allocation*

The Great Depression in the United States is an example of a large deviation from the neoclassical growth theory that is not accounted for by variations in TFP. In 1939 hours worked per adult were still 23 per cent below what they were in 1929, the year prior to the start of the Depression. During this ten year period output per hour increased by about 10 per cent, which is only a little below the historical average. The question is why employment did not return to its 1929 level. The only candidate for an answer is policy that changed the nature of the game being played by the economic actors.

### 12.2.4 International business cycles

Backus, Kehoe and Kydland (1992) use theory to construct a model economy to derive the implications of growth theory for international business cycles. They assume, among other things, that there is a full set of state-contingent commodities. They find that observations deviate from the predictions of this theory in important respects. First, consumption is less correlated between countries than theory predicts. Second, investment and labor supply are too negatively correlated with output.

Their findings led Baxter and Crucini (1995) and Kollmann (1996) to restrict international asset markets to just borrowing and lending. With this restriction, the model economy better mimics reality. But, the question of why there is this restriction is left unanswered. This led Kehoe and Perri (forthcoming) to endogenize these debt contracts. They consider a model economy with limited enforcement of international contractual constraints.

### 12.2.5 Introducing contractual constraints using modern contract theory

Cooley and Quadrini (1998) find that if firms face debt constraints, monetary policy has quantitatively important welfare effects. The economy mimics reality on a number of dimensions suggesting that indeed firms are debt constrained. Here, I think, there is a need to use modern contract theory to endogenize these constraints. The issue of what is a firm, which in Arrow–Debreu theory is just a technology set, must be addressed. Here I conjecture that modeling a firm as a coalition or club is the way to proceed. Useful tools include the recursive dynamic contract theory of Thomas and Worrall (1990), Marcet and Marimon (1992), and Atkeson and Lucas (1992,1995). Other useful tools might be general equilibrium theory with clubs as in Cole and Prescott (1997) and financial intermediary coalitions as in Boyd and Prescott (1986).

Carlstrom and Fuerst (1997), in an interesting paper, deduce some quantitative implications of these informational constraints on contracting. Alvarez and Jermann (2000) quantitatively explore the implications of limited enforcement for asset prices in an applied general equilibrium study of asset pricing.

### 12.2.6 Introducing plant irreversible investment

Marcelo Veracierto (1998) develops a method for introducing irreversible investment at the plant level into equilibrium business cycle models. The author's computational strategy makes the analysis of the class of $(S, s)$ economies fully tractable. He finds that introducing this feature has no effects on aggregate business cycle dynamics.

### *12.2.7 Computing equilibrium when a distribution is part of the state variable*

There are many issues that require a model with a distribution as a state variable. Existing tools for analyzing such model economies are limited although tools are improving. The problem is the well know curse of dimensionality in recursive analyses. Veracierto (1998) finesses this problem with his approach. Krusell and Smith (1998) have developed methods to find an approximate solution that turns out to be a remarkably good approximation in some applications.

### *12.2.8 Role of financial costly financial intermediation in business cycle*

Large amounts of resources are used in financial intermediation. In the United States resources used are approximately 9 per cent of GNP. The average spread between average borrowing and average lending rates is about 5 per cent. Further there is a lot of borrowing and lending between households that is intermediated through the business sector. In the United States, governments have nominal liabilities equal to about one annual GNP.

### *12.2.9 The role of varying number of shifts that plants are operated*

A margin of adjustment that has not been introduced into any applied general equilibrium analysis of business cycles is the option to vary the number of shifts a plant is operated. This is an important margin of adjustment in the automobile industry.

## Acknowledgement

Prepared for the International School of Economic Research's XI Workshop, 'Cycles, Growth and Technological Change.' 29 June 1998, to 7 July 1998. Certosa di Pontignano-Siena (Italy). I thank Dirk Krueger and Jessica Tjornhom for useful comments and assistance in preparing the list of references. The views expressed herein are those of the author and not necessarily those of the Federal Reserve Bank of Minneapolis or the Federal Reserve System.

## Notes

1. See Gollin (1997) for the cross-country numbers. He uses the Kravis' (1959) economy-wide assumption for assigning proprietor's income and indirect business taxes to capital and labor.
2. See Chari, Kehoe and McGrattan (2000) for a critical evaluations of mechanisms to propagate the effects of monetary shocks.

## References

Alvarez, F. and U. Jermann (2000) ‘Efficiency, equlbrium and asset pricing with risk of default’, *Econometrica*, 68, (4), 775–98.

Atkeson, A. and R. Lucas (1992) ‘On efficient distribution with private information, *Review of Economic Studies*, 59, (3), 427-53.

Atkeson, A. and R. Lucas (1995) ‘Efficiency and equality in a simple model of efficient unemployment insurance’, *Journal of Economic Theory*, 66, (1), 64-88.

Backus, D., P. Kerhoe and F. Kydland (1992) ‘International Real Business Cycles’, *Journal of Political Economy* 101, 745-75

Baxter, M. and M. Crucini (1995) ‘Business cycles and the asset structure of foreign trade’, *International Economic Review*, 36, (4), 821-54.

Boyd, J. and E. Prescott (1986) ‘Financial intermediary coalitions’, *Journal of Economic Theory*, 38, (2), 211-32.

Bresnahan, T. and V. Ramey (1993) ‘Segment shifts and capacity utilization in the US automobile industry’, *American Economic Review*, 83, (2), 213-18.

Brock, W. and L. Mirman (1972) ‘Optimal economic growth and uncertainty, the discounted case’, *Journal of Economic Theory*, 4, (3), 479-513.

Burns, A.F. and W.C. Mitchell (1946) *Measuring Business Cycles*, New York: National Bureau of Economic Research.

Burnside, C., M. Eichenbaum and S. Rebelo (1993) ‘Labour Hoarding and the Business Cycle’, *Journal of Political Economy* 101, 245–73.

Carlstrom, C. and T. Fuerst (1997) ‘Agency costs, net worth, and business fluctuations: a computable general equilibrium analysis’, *American Economic Review*, 87, (5), 893-910.

Cass, D. (1965) ‘Optimum growth in an aggregative model of capital accumulation’, *Review of Economic Studies*, 32, (3), 233-40.

Chari, V.V., P. Kehoe and E. McGrattan (1998) Monetary shocks and real exchange rates in sticky price models of international business cycles, Federal Reserve Bank of Minneapolis Staff Report 223.

Chari, V.V., P. Kehoe and E. McGrattan (2000) ‘Can the contract multiplier solve the persistence problem?’, *Econometrica*, 68, (5), 1151–79.

Christiano, L. and M. Eichenbaum (1995) ‘Liquidity effects, monetary policy, and the business cycle’, *Journal of Money, Credit and Banking*, 27, (4), 1113-36.

Christiano, L. and M. Eichenbaum (1997) ‘Sticky price and limited participation models of money, a comparison’, *European Economic Review*, 41, (6), 1201-49.

Cole, H. and E. Prescott (1997) ‘Valuation equilibrium with clubs’, *Journal of Economic Theory*, 74, (1), 19-39.

Cooley, T., G. Hansen, and E. Prescott (1995) ‘Equilibrium business cycle with idle resources and variable capacity utilization’, *Economic Theory*, 6, (1), 35-49.

Cooley, T. and V. Quadrini (1998) ‘Monetary policy and the financial decisions of firms’, *mimeo*.

Devereux, M., A. Head, and B. Lapham, (1996) ‘Aggregate fluctuations with increasing returns to specialization and scale’, *Journal of Economic Dynamics and Control*, 20, (4), 627-56.

Diamond, P. (1965) ‘National debt in a neoclassical growth model’, *American Economic Review*, 55, (5), 1126-50.

Diáz-Giménez, J., E.C. Prescott, T. Fitzgerald and F. Alvarez (1992) ‘Banking in Computable General Equilibrium Economies’, *Journal of Economics Dynamics and Control* 16, 533–59.

Fagnart, J., O. Licandro, and F. Portier (1999) 'Firm heterogeneity, capacity utilization and the business cycle', *REview of Economic Dynamics*, 2, (2), 433–55.

Fitzgerald, T. (1996) 'Reducing working hours', *Economic Review, Federal Reserve Bank of Cleveland*, 32, (4), 13-22.

Fitzgerald, T. (1998) 'Work schedules, wages, and employment in a general equilibrium model with team production', forthcoming in *Review of Economic Development*.

Freeman, S. and F. Kydland (2000) 'Monetary aggregates and output', *American Economic Review*, 90, (5), 1125–35.

Friedman, M. and A.J. Schwartz (1963) *A Monetary History of the United States, 1867–1960*, Princeton: Princeton University Press.

Fuerst, T.S. (1992) 'Liquidity, Loanable Funds and Real Activity', *Journal of Monetary Economics*, 29, 3–24.

Gollin, D. (forthcoming) 'Getting income shares right', *Journal of Political Economy*.

Greenwood, J., Z. Hercowitz and G.W. Huffman (1988) 'Investment, capacity utilization and the real business cycle', *American Economic Review* 78, 402–17.

Hansen, G.D. (1985) 'Indivisible labor and the business cycle', *Journal of Monetary Economics*, 16, 309-27.

Hart, O. (1979) 'Monopolistic competition in a large economy with differentiated commodities', *Review of Economic Studies*, 46, (1), 1-30.

Hodrick, R. and E. Prescott (1980 and 1997) 'Postwar US business cycles: an empirical investigation', *Journal of Money, Credit, and Banking*, 29, (1), 1-16 (1997), Discussion Paper 451, Northwestern University (1980).

Hornstein, A. and J. Praschnik (1997) 'Intermediate inputs and sectoral comovement in the business cycle', *Journal of Monetary Economics*, 40, (3), 573-95.

Hornstein, A. and E. Prescott (1993) 'The plant and the firm in general equilibrium theory', in R. Becke, M. Boldrin, R. Jones and W. Thompson (eds) *General Equilibrium and Growth: The Legacy of Lionel McKenzie*, New York: Academic Press.

Ireland, P. (1995) 'Endogenous financial innovation and the demand for money', *Journal of Money, Credit and Banking*, 27, (1), 107-23.

Johnson, H. (1960) 'The cost of protection and the scientific tariff', *The Journal of Political Economy*, 68, (4), 327-45.

Jones, L. (1984) 'A competitive model of commodity differentiation', *Econometrica*, 52, (2), 507-30.

Kehoe, T. and D. Levine (1993) 'Debt constrained asset markets', *Review of Economic Studies*, 60, (4), 865-88.

Kehoe, P. and F. Perri 'International business cycles with endogenous market incompleteness', *Econometrica* (forthcoming).

Kehoe, T. and E. Prescott (1995) 'Introduction to the symposium, the discipline of applied general equilibrium', *Economic Theory*, 6, (1), 1-11.

Kollmann, R. (1996) 'Incomplete asset markets and the cross-country consumption correlation puzzle', *Journal of Economic Dynamics and Control*, 20, (5), 945-61.

Koopmans, T. (1965) 'On the concept of optimal growth', *The Econometric Approach to Development Planning*, Rand McNally, Chicago.

Kravis, I. (1959) 'Relative income shares in fact and theory', *American Economic Review*, 49, (5), 917-49.

Krusell, P. and A. Smith (1998) 'Income and wealth heterogeneity in the macroeconomy', *Journal of Political Economy*, 106, (5), 867–96.

Kuznets, S. (1946a) *National Income: A Summary of Findings*, National Bureau of Economic Research (NBER), New York.

Kuznets, S. (1946b) *National Product Since 1869*, National Bureau of Economic Research (NBER), New York.

Kydland, F. and E. Prescott (1982) 'Time to build and aggregate fluctuations', *Econometrica*, 50, (6), 1345-70.

Kydland, F. and E. Prescott (1991) 'Hours and employment variation in business cycle theory', *Economic Theory*, 1, 63–81.

Kydland, F. and E. Prescott (1996) 'The computational experiment: an econometric tool', *Journal of Economic Perspectives*, 10, (1), 69-85.

Lucas, R. (1977) 'Understanding business cycles', *Journal of Monetary Economics*, 5, (0), 7-29.

Lucas, R. (1980) 'Equilibrium in a pure currency economy', *Economic Inquiry*, 18, (2), 203-20.

Lucas, R. (1981) *Studies in Business-Cycle Theory*, MIT Press, Cambridge, MA.

Lucas, R. (1988) 'Money demand: a quantitative review', in K. Brumer and B. McCallum (eds) *Money, Cycles and Exchange Rates: Essays in Honor of Alan Meltzer*, Carnegie-Rochester Conference Series on Public Policy, vol. 29, North Holland, Amsterdam.

Lucas, R. and N. Stokey (1987) 'Money and interest in a cash-in-advance economy', *Econometrica* 55, 491–514.

Marcet, A. and R. Marimon (1992) 'Communication and growth', *Journal of Economic Theory*, 58, 219–49.

Mas-Colell, A. (1975) 'A model of equilibrium with differentiated commodities', *Journal of Mathematical Economics*, 2, 263-96.

McCallum, B. and M. Goodfriend (1987) 'Money, theoretical analysis of the demand for money', *NBER Working Paper No. 2157*.

Prescott, E. (1986) 'Theory ahead of business cycle measurement', Carnegie-Rochester Conference Series on Public Policy, vol 24. Reprinted in *Federal Reserve Bank of Minneapolis Quarterly Review* 10, 9–22.

Prescott, E. and R. Townsend (1984) 'Pareto optima and competitive equilibria with adverse selection and moral hazard', *Econometrica*, 52, (1), 21-45.

Saving, T. (1971) 'Transaction costs and the demand for money', *American Economic Review*, 61, (3), 407-20.

Scheinkman, J. and L. Weiss (1986) 'Borrowing constraints and aggregate economic activity', *Econometrica*, 54, (1), 23-45.

Schreft, S. (1992) 'Transaction costs and the use of cash and credit', *Economic Theory*, 2, (2), 283-96.

Schumpeter, J. (1954) *A History of Economic Analysis*, Oxford University Press, New York.

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), 367-90.

Townsend, R. M. (1979) 'Optimal contracts and competitive markets with costly state verification', *Journal of Economic Theory*, 21, (2), 265-93.

Veracierto, M. (1998) 'Plant level irreversible investment and equilibrium business cycles', *mimeo*.

# 13 Complexity-based methods in cycles and growth

## Any potential value-added?

*William A. Brock*

## 13.1 Introduction

This chapter is written in the style of both a brief survey and a research proposal in order to present a set of analytical and statistical tools, some of which are related to what might be called ‘complexity theory’, and to explore whether these tools have anything to contribute to the study of cycles, growth, and structural change. The reader is warned at the outset that much of what is said here is extremely speculative and may end up being a research dead end. Furthermore, this article will keep mathematics and technical details to a minimum in order to attempt to reach a wide audience.

I shall give a lot of references to the literature for details. I apologize in advance for the literary license with subtle ideas that I take here. Defining whether an analytical, theoretical, or statistical tool is ‘complexity-based’ or a research methodology is ‘complexity-based’ is hard to do within the scope of a short article. So we shall just ask the reader to take a look at ‘complexity-based’ references to see what the word ‘complexity’ refers to. Many ‘complexity-based’ research groups have been operating for a long time.

There are, for example, the Brussels School (cf. Day and Chen (1993) for a sampling from the Brussels/Austin school as well as others), Day’s two volume work (Day (1997) which covers not only the work of himself and his co-workers but also work of many others), Rosser’s book ((1991), which gives an exceptionally broad based coverage of the work of different research groups in this area), and the Stuttgart School (cf. Weidlich (1991)).

The current article will draw on my association with the Santa Fe Institute (SFI). The books, Anderson, Arrow, and Pines (1988), (SFI (I)), and Arthur, Durlauf, and Lane (1997), (SFI (II)) give nice descriptions of ‘complexity theoretic’ approaches to economics from the ‘SFI’ point of view. For example, let us quote from the introduction to SFI (II):

> But just what is the complexity perspective in economics? That is not an easy question to answer. Its meaning is still very much under construction. . . . To describe the complexity approach, we begin by pointing out six features of

> the economy that together present difficulties for the traditional mathematics used in economics.

We list them as follows: (i) dispersed interaction; (ii) no global controller; (iii) cross-cutting hierarchical organization; (iv) continual adaptation; (v) perpetual novelty; (vi) out-of-equilibrium dynamics. Notice that all six being present together is what prevents one from treating them by simply re-interpreting the usual Arrow–Debreu–McKenzie apparatus of general equilibrium theory with enlarged state spaces of date-event contingent goods and the like with disequilibrium dynamics grafted onto this apparatus. For example, Norman Packard argued in his piece in Anderson, Arrow, and Pines (1988) that the state space's dimensionality is constantly changing so we cannot use conventional dynamical systems theory to analyze the dynamics generated by his complex system.

Readers can be quite assured that the piece they are reading is 'complexity-based' if they encounter the following terms and ideas: (a) path dependence; (b) self-organized criticality; (c) edge of chaos; (d) power law scaling; (e) renormalization group; (f) fractals and other types of self similarity; (g) genetic algorithms, emergent computation, (h) adaptive neural networks, complex adaptive systems; (i) chaos theory, embedding theorems, correlation integrals; (j) interacting particle systems, statistical mechanics, mean field theory, non-ergodic systems, breakdown of the law of large numbers.

My paper, (1999), compares and contrasts, for applications to economics, the SFI (I,II) version of complexity approaches to economics with another closely related approach which emphasizes a hierarchy of dynamical systems at temporal and 'spatial' scales and uses bifurcation theory from dynamical systems theory.

The latter research style is common in ecology. Rather than repeat what has been stated already in SFI (I,II) and Brock (2000), we shall just sketch the needed highlights here and refer to these sources for details. The econometric and statistical tools reviewed by Brock and Durlauf (1999) will be used to propose a different twist to the problem of detecting and measuring the 'endogeneity' in 'endogenous' growth theories. I shall also outline some proposed research strategies to help narrow the dispute between those who prefer rational expectations models of cycles and growth and those who prefer 'boundedly rational' approaches.

## 13.2 Total factor productivity viewed with 'complexity-based' tools

Before we are in a position to discuss whether there is any potential value-added of complexity-based tools to students of economic growth and cycles, we must spend some time reviewing 'conventional' work. I shall use optimal timing theory to add a tiny bit of novelty, perhaps, during this part of the review.

Let us use a discussion of Total Factor Productivity (TFP) dynamics as the central organizing idea of this part of the chapter. A related 'complexity-based' review of the dynamics of technical change, i.e. TFP dynamics, is Dosi (2000). My objective will be to isolate the value-added of complexity-based approaches to econometrics of TFP dynamics. For example, the notion of path dependence raises the issue of testing for 'true state dependence' versus 'spurious state dependence' as in Heckman and Singer (1985) and measuring the relative strengths of each using actual data. The notion of 'lock-in' raises the issue of alternative stable states, especially Pareto-ranked alternative stable states, which, in turn, raises the econometric issue of testing and measuring such effects. It also raises the issue of isolating the obstruction to economic agents or coalitions of economic agents to picking up 'big bills on the sidewalk'. (Olson (1996)). This raises the econometric issue of isolating and gathering the kinds of data we need to identify these obstructions and to measure the size of these obstructions. Indeed one might say it raises the issue of how one might test whether the current state of an economy is Pareto Optimal or not. Turn now to TFP.

It seems fair to say that one of the key issues discussed by several papers presented at the Siena School in July, 1998 was finding good explanations for the tremendous differences in Total Factor Productivity (TFP) growth across countries as well as standards of living measured by, for example, Gross Domestic Product (GDP) per capita.

Baumol's paper (Baumol (2001, this volume)) discussed relative TFP growth over long time scales of centuries and stressed institutions which greased the competitive process which generated pressure for each business to innovate before some other business out innovates it and puts it out of business. He also stressed large spillovers that were not captured by the innovators. He used the illustration problem of the optimal time to buy a computer when the cost was falling and the usefulness was rising in order to focus attention on economic forces that speed up adoption of innovations. In the spirit of Baumol, I shall use optimal timing theory to shed light on some economic forces that influence TFP growth. This will motivate a later discussion of the econometrics of duration models with social interactions following Brock and Durlauf (1999) and Heckman and Singer (1985). Surveys of growth theory, both empirical and theoretical, such as Barro and Sala-i-Martin (1995) attack the TFP question at several scales of time ranging from centuries to decades, especially post-World War II performance. Barro's recent book (Barro (1997)) surveys his well known work. The paper of Durlauf and Quah (1999) gives us a tour of the whole field as well as a critical review which stresses basic econometric identification issues. A major theme that emerges from this aggregative literature is the difficulty of empirically differentiating between modifications of 'Solovian' type theories and the 'new endogenous growth' theories, by conditioning on regressors that attempt to capture the quality of infrastructure, human capital, and incentive structures (embodied, for example, in the quality of institutions) Indeed Durlauf and Quah (1999) show how difficult it is to use available data to even distinguish between a broadly interpreted Solow type model with 'mechanical' savings behavior such as Mankiw, Romer and Weil with two types of capital from an optimizing model of such savings behavior. I.e. it appears to be difficult to use available data to distinguish between growth models that are 'Solovian modifications' with 'boundedly rational expectations' from models which are modifications of the rational expectations 'forward looking' neoclassical models.

Turnovsky's book (1995) has reviewed generalizations of the basic intertemporal general equilibrium growth models to include policy induced distortions such as taxes at both the corporate level and the personal level, inflationary finance, debt finance, and others. This raises the issue of how to use these recently developed growth models to design econometric methods that help detect the 'jump variable' effects associated with anticipated (via forward looking rational expectations type behavior) policy changes (cf. Solomou (1998) for a critical review in applications to cycles and growth analysis). A major issue is how to empirically distinguish lagged effects such as adjustment cost effects in rational expectations models from lagged effects due to backwards looking expectations (e.g. boundedly rational expectations as in Sargent (1993)).

The usual approaches in rational expectations econometrics include, for example: (i) formulate and test cross equation restrictions (cf. Hansen and Sargent (1991)); (ii) formulate and test for the small real impact of, for example, 'anticipated money' in contrast to the predicted larger real effect due to 'un-anticipated money' under rational expectations in contrast to some types of boundedly rational expectations. See Solomou (1998) for a review and critique of attempts to use (ii) in cycle analysis.

A 'nested' approach applied by Baak (1999) might be adaptable to the problem of distinguishing boundedly rational expectations from rational expectations in cycles and growth models. Here one writes down a general model where rational expectations are costly and compete against other types of expectations which are more boundedly rational and also 'cheaper to purchase'. One then estimates this model and tests the null hypothesis whether the 'extra parameters' that appear in the general equilibrium equations are 'significant.' This is one way to assess whether the extra free parameters brought in by boundedly rational theorists cover their 'cost' in terms of extra predictive and explanatory power. More will be said about this kind of procedure later.

Empirical problems loom even larger when one attempts to econometrically identify and estimate the 'spillover' effects (which have 'social multipliers') which were stressed in the early Romer models and in the Lucas/Uzawa models reviewed in Barro and Sala-i-Martin (1995). Recall that the early Romer models stressed noninternalized spillovers from production activities and the Lucas/Uzawa type models stressed noninternalized spillovers from human capital. We shall discuss this last econometric identification problem using econometric work on identifying 'true social interactions' from 'spurious social interactions' using tools reviewed by Brock and Durlauf (1999) and Manski's paper in SFI (II).

Prescott's Klein Lecture (Prescott (1998)) attacks the same issue of explaining the huge disparity of living standards across countries and argues that we need a theory of TFP. He outlines some basic historical facts about TFP performance across countries and within industries and poses strong challenges to the literature to date to explain these facts. Hahn's paper (2001, this volume) argues that so called 'endogenous growth theories' are 'exogenous' after all. I.e. much is left unexplained rather than explained. It may be useful to get started by thinking about cycles, growth, and structural change in terms of a hierarchy of time scales because different tools, methods of analysis and explanations may be needed at different time scales. Day and Pavlov's (2001, this volume), Baumol's (2001, this volume), and Barro's (1997) work stressed the determinants of economic progress on the time scale of millenia, centuries, and decades respectively. Business cycle workers concentrate on yearly frequencies because business cycles have a modal length of 3–5 years. Call these time scales, very slow, slow, medium, and fast.

Useful analytical hierarchies of time scales are common in economics. For example, in financial analysis the typical hierarchy of time scales ranges from business cycle scale as slowest to minute by minute scale (tic by tic) as fastest. Furthermore in subjects where the phenomena under scrutiny have a 'spatial' relationship (where 'space' is broadly interpeted) as well as a temporal relationship it is useful to introduce a spatial 'hierarchy' of scales from largest to smallest.

This theme of a hierarchy of time scales is used in Böhm and Punzo (2001, this volume) where they stress that variables at a slower scale serve as approximate comparative statics parameters for variables moving at the next faster time scale in the hierarchy. Indeed the slower variables may pass through a bifurcation point in the Böhm and Punzo work.

The work of the Resilience Network, hereafter, 'Rnet', discussed in Brock (2000) stresses the use of a hierarchy of time scales as a useful device in organizing an analysis of dynamic phenomena in general. See Gunderson, Holling, and Light (1995) for a sample of Rnet approaches to ecosystem management. Rnet models stress the potential presence of alternative stable states and adducing empirical evidence for or against the alternative stable state hypothesis. This emphasis is similar to the nonconvex models of economic growth of Azariadis and Drazen, Galor and Ziera, Murphy, Shleifer, and Vishny which are discussed by Durlauf and Quah (1999) (DQ) from the perspective of adducing empirical evidence for clustering and divergence in growth performance across countries.

Rnet models, however, stress the existence of slow moving background dynamics which are difficult for observers to detect and measure but which nevertheless impact the ability of the 'primary dynamics' to absorb shocks and revert to steady state after such shocks. In DQ (1999) continuous time dynamical language, the stable eigenvalue (a pair of eigenvalues with negative real parts if complex) which measures how fast the system reverts back to its steady state path after a shock is impacted by a state variable driven by a slow moving dynamics (which is fed by the primary dynamics).

For example the stable eigenvalue may be gradually moved closer to zero by the slow moving dynamics, but this is not apparent to observers, (both to scientists outside the system and to actors inside the system) until a really large shock hits the system which makes it apparent to both outside and inside observers that the system's capacity to absorb shocks has been compromised. For an example of this type of modelling of ecosystems that might be adapted to growth modeling see Brock (200) and, especially, Gunderson *et al*. (1995).

In some phenomena such as climate dynamics there will be a rough positive relationship between the time scale and spatial scale of an activity. A quantitative depiction of this relationship, which plots some measure of relative strength or relative frequency of the activity in that portion of space/time relative to other portions of space/time such as spectral power on a temporal axis and on a spatial axis in an attempt to isolate 'spectral mountains', called a Stommel diagram, is sometimes used in other sciences but I have not seen it used in economics. See Gunderson *et al*. (1995) for some examples of Stommel diagrams. More will be said about this later when we discuss 'spatial' growth models (e.g. Durlauf's lattice-growth model and its relatives discussed in Durlauf's piece in Anderson, Durlauf and Lane (1997)), and 'sandpile' models (See, e.g. Scheinkman and Woodford (1994), Krugman (1996), and references to Bak and others).

For example, on the slow time scale of centuries, Baumol's (2001, this volume) work stressed the determinants of economic progress over centuries and asked the question why, at this time scale, have the Western nations progressed so far in improving the standards of living of their peoples relative to many others. He stressed the institutional environments in these countries which puts such strong pressure on each business to innovate under the constant threat of being put out of business by one who does. He stresses the construction of these institutions and the degree of spillovers that accrue to the population at large.

This theory is related to the 'Schumpeterian' family of theories, which appear, for the most part, to be designed to explain economic progress at the medium scale. See, for example, writers such as Aghion and Howitt, and Jovanovic in Kreps and Wallis (1997), Jovanovic and Lach (1997), as well as Iwai (2001, this volume) and Aghion and Howitt's recent book (1998). I shall stress incentives to adopt existing technology in the style of Jovanovic's piece in Kreps and Wallis, Harberger (1998), Prescott (1998), rather than the grand incentive to innovate new ideas. The grand incentives and Schumpterian waves of innovations can be thought of as forces on a slower time scale than we wish to discuss here. See, for example, Solomou (1998) for a general empirical and historical analysis on these slower scales in comparison to business cycle analysis on faster time scales. I adopt a perspective much like Solomou here. For example, he writes (1998, p. 119)),

> The many changes in cyclical features reflect, among other things, changes to economic structures, policy frameworks and behavioural patterns. Given that these features are in a state of flux we are unlikely to observe a universal business cycle structure that can be understood by explaining the features of particular periods. . . . We cannot hope to predict business cycles, but we can hope to understand the foundations for making conditional predictions.

The analysis, to be conducted below, of the impact of five industrial organizational and regulatory arrangements, I1–I5, on the rate of adoption of TFP enhancing activities is in the spirit of attempting to understand some 'foundations for making conditional predictions' about the impacts of different 'economic structures, policy frameworks, and behavioural patterns'. One might also imagine that innovation activity at the Schumpetarian level of 'high imagination' would be much more difficult to do econometrics on (how do you predict when the next Crick/Watson type of finding will set off a biotech type wave?) than predicting what will happen if one restructures a state-owned enterprise so that top executives are paid a bonus based upon a measure of TFP growth rather than flat government salaries. We doubt, anyway, that the relative performance of North Korea relative to South Korea is due to a lack of Crick and Watsons in North Korea relative to South Korea. A major challenge posed by writers like Olson (1996) and Prescott (1998) is this. Why doesn't the set of nonadvanced nations, especially the poor ones, of the world simply copy and adopt the huge existing stock of technology built up by the advanced nations and give their peoples a 'free' one-shot jump in their standard of living up to the level of the advanced nations? Here we might think of the 'stock of technology' broadly enough to include the stock of organizational 'capital' and the stock of institutional 'rules-of-the-game' experience that is embodied in the collective experience and history of the advanced nations. What explanation for this 'adoption gap' can economists give that advances the discussion beyond that given by a perceptive tourist? Let us peer inside the black box of TFP growth with a slightly different emphasis than the recent literature by using some tools from optimal timing theory including the Feynman–Kac formula. The approach will draw attention to variability of the stock of 'latent options' on both the quality improvement side and finding new markets side of growth as well as the cost reduction side of growth. For example, the idea of latent options is to shed some potentially slightly different light on the economic incentives to pay a cost in order to obtain what Harberger (1998) calls a Real Cost Reduction (RCR). Our use of optimal timing theory to expose economic forces that drive TFP dynamics motivates the proposed use of the econometrics of hazard function estimation to be discussed below.

The kind of cost we have in mind is more related to that paid by a business executive in dealing with a recalcitrant middle management and labor force to change their past set of habits and work routines in order to adopt the new way of production. Or, at a more personal level, think of the problem of finding the optimal time to switching to a new word processing program as the hassle and learning cost of switching falls at some rate (perhaps because newer, more user-friendly versions are appearing and more people in your office are switching so you can learn from them) while the benefit to switching rises at some rate. In any event, as Harberger stresses, this kind of RCR takes place in 1001 ways.

To put it more bluntly we attempt to shed light on forces that remove incentives to indulge in gross inefficiencies such as over-manning an enterprise, over-use of capital, ignoring obvious markets, catering to factions such as labor and management by using the enterprise as a comfort and security vehicle for these factors instead of serving the customers for whom the enterprise was set up to serve. These types of obviously inefficient behaviors are typically associated with regulated industries, government departments, protected industries, and the like. These kinds of perceived gross inefficiencies gave rise to reforms such as the 'corporatization reforms' in New Zealand as well as 'privatization' reforms world wide. See Evans *et al*. (1996) for a careful review of the restructuring experience in New Zealand. See Olson (1996) for a dramatic argument that 'economic performance is determined mostly by the structure of incentives – and that it is mainly national borders that mark the boundaries of different structures of incentives' (Olson, 1996, p. 22).

For a dramatic review of problems in transition economies see the Symposium on Transition from Socialism in the same issue of JEP as Olson's lecture. See also McMillan and Dewatripont and Roland's pieces in Kreps and Wallis (1997) for what might be called the '1001 complementarities' of institutional and experiential infrastructure that are needed to align the structure of incentives of the agents within a transition economy in the direction of the public interest.

Notice that both Prescott (1998) and Jovanovic's piece in Kreps and Wallis (1997) stress factors closely related to the structure of incentives when they stress how successful resistence to adopting pieces of the huge already existing stock of usuable technology (not only technical but also administrative) might explain a lot of the current gap in standards of living across countries. We do not wish to focus on the costs of developing new technologies like biotech because that is well covered in the literature, although the idea of latent options stressed in this article applies here too. See especially Romer's website (Romer (1998)) for a vivid discussion of the power packed by potential innovations for economic progress.

We also abstract away from negative externalities such as environmental degradation due to adoption of the RCR. I.e. we are assuming that incentives are aligned to the social accounting system which ensures that all social costs are borne on private account. I.e. Social RCRs are identical to Private RCRs. We are focusing on the general problem of harnessing the Schumpeterian engine of economic progress for the social interest. The very important problem of incentive design to redirect the Engine from finding private RCRs that advance the private interest and that off-load costs onto the rest of society towards finding social RCRs that advance the social interest where all cost-causers bear their own costs is left for another time. The approach taken here will be to write out a traditional structure–conduct–performance outline of the impact on the rate of TFP growth due to the speed of adoption of RCRs using existing technology. I hasten to add that, except possibly for a slightly different approach to the idea of latent options, I am adding nothing new to a very well-worked literature. However some of the 'tool scenery' may be interesting and may be useful to readers in their own research. After, hopefully shedding some light on quantifying the strengths of carrot and stick forces to innovate under different industrial structures and different institutions and, perhaps, adding a little bit to the well developed literature in this area, I wish to turn to the problem of identification of spillovers that catalyze growth and positive feedback forces in the 'growth regressions' literature reviewed by Durlauf and Quah (1999). This discussion will use some econometric tools developed in Brock and Durlauf (1999) and some theoretical 'complexity' tools developed in Brock (1993) and Brock and Durlauf (1995).

The treatment of the well-worked area of innovation and adoption theory will be brief. Optimal stopping theory and the Feynman–Kac formula (e.g. Duffie (1988)) will be used to add a little material on the interaction of variance and industrial structure that might shed a bit of light on forces behind TFP growth that the existing literature has not quite cleaned up. Before we begin this part let us consider growth regression analysis and state the basic equation used in such analysis from Durlauf and Quah (1998), which we denote by (DQ.19),

$$\log y(t) = \log(b) + \log(A(0)) + gt + [\log(y(0)/b) + \log(A(0))]\exp(lt), \qquad \text{(DQ.19)}$$

where $y(t)$ denotes observable per capita income, $A(t)$ denotes the 'level of technology' at date $t$, $\mathrm{d}\log(A(t))/\mathrm{d}t = g$, $b$ and $l$ are parameters that depend upon other economic parameters such as the rate of population growth, the rate of depreciation on capital stock, savings rate, and the like.

DQ show how (DQ.19) emerges from the main received growth theories and they use it to organize their discussion of forces that shape the 'convergence component' (the term involving $\exp(lt)$) and the 'levels component (the rest of the right-hand side).' They point out that since $\log(b), \log(A(0))$ are unobserved just about any pattern of 'cross-country growth and convergence is consistent with the model.'

Thus we need to model the forces that shape $\mathrm{d}log(A(t))/\mathrm{d}t = g, A(0), b$, and $l$. Hence, while reading the part below, readers might view it as a discussion of some economic forces that shape $g = \mathrm{d}\log(A(t))/\mathrm{d}t$. Readers should imagine that they have data on a cross section of countries where they are estimating a system like (DQ.19) with one equation like (DQ.19) for each country. Theory like that presented below will be used to suggest regressors to be inserted into the RHS of (DQ.19). Economic forces that lead to diffusion of technologies across different countries such as international trade lead to cross equation dependence for systems like (DQ.19). More will be said about estimation of such systems later.

### *13.2.1 Optimal stopping tools and exploitation of options*

I shall draw on optimal stopping theory to compute the optimal time of introduction of an innovation or project which costs $F$ at time of introduction and generates a flow $S(t)$ of net benefits from that time on. The introduction cost $F$ will be widely interpreted so that it can stand for costs of overcoming employee resistence to a new technique following Prescott's discussion (1998). A complete benefit/cost welfare analysis might include the discomfort costs of adapting to the new innovation that are borne by the businesses executives, middle management, and workers. More will be said about this below. The results will be organized in a structure–conduct–performance framework so familiar from theoretical industrial organization analysis, but carried out from the vantage point of optimal stopping theory with the goal of contributing to understanding of the forces that shape the rate of growth of TFP. This analysis exploits elements of the work of Brock, Rothschild and Stiglitz (1989) and their references to the works of Brock, Miller, Scheinkman, and Ye.

I will use these qualitative results to discuss probable impacts on the time evolution of the TFP quantity, $\log(A(t))$, in the Durlauf and Quah (1999) survey of the empirical growth literature. First I shall analyze the simplest possible problem under five industrial organizations.

- I1 denotes the social interest.
- I2 denotes a monopolist who cannot price discriminate and, hence, cannot extract out the area under the demand curve at each point in time.
- I3 denotes competition between two firms and the first to build will have a monopoly from that point on.
- I4 is same as I3 but the monopolist will have the ability to price discriminate.
- I5 Is the general problem where the incentive system of the management puts different weights on the consumer benefit side and the adoption side of the problem. This may be due to distortions induced by regulatory systems, the tax system, whether the enterprise is a government department, a partial public-private enterprise, etc.

Although we placed I5 at the end, it may be the most important of all. Many economic entities operate under incentive systems placed upon them that 'tax' heavily the gains to potential productivity enhancing changes in operations but do not give full 'tax-offset,' i.e. 'tax deduction,' for the costs, both personal and corporate of carrying out those changes. For example the management may realize that their enterprise is inefficient but it does not want to bear the personal cost of having to fire people. If the tax system captures 50 per cent of the profit, the management's contract gives a bonus of 1 per cent of the after tax profit, and the personal income tax takes half of this, while there is no offset for the personal cost of 'taking the heat' of firing the excess labor, it is not surprising that the enterprise will be over-manned. See Gibbons' piece in Kreps and Wallis (1997) for a review of work on incentive design in organizations. Turn now to the simplest problem for analysis. Let's start with the simplest problem which I shall call the DATRAN problem after one of my articles in Evans (1983, Chapter 8)). This is the problem of identifying a new market whose profit flow is growing at some fixed positive rate and building a facility that costs $F$ to capture it. Properly measured TFP will stress the development of new goods and new markets as emphasized by DeLong where he reviews the proper measurement of economic progress at his website (1998).

Later we shall discuss the cost side where there is technology available which is reducing unit cost at some rate '$c$' but it costs $F$ to adopt it. The analysis of both problems is similar. Here is a prototype example problem. When should society build a digital data transmission network which will cost a fixed setup cost $F$, will serve the entire demand, and where the social value of the net benefit per period generated by the network $S(t)$ is growing at constant rate $a$? To be more precise let demand at date $t$ be given by $D(q)\exp(at)$, gross benefit be the area under $D(\cdot)\exp(at)$ up to $q$, call this $B(q)\exp(at)$, let unit cost be constant in $q$ at all dates $t$ and be given by $c\exp(at)$. Letting gross benefit and cost grow at the same rate makes the mathematics simple. This problem is the problem of when to build a facility to capture a new market. Alternatively it may be viewed as the problem of selecting the optimal time to adopt an RCR which yields a gain which is growing at rate '$a$' but the cost to adopting is $F$. The RCR adoption problem where cost $F$ is falling at some rate has a very similar structure to this problem, as we shall see later. Both problems concern finding the optimal time to take advantage of a latent opportunity. The net benefit, $S(t) = \exp(at)[\max\{B(q) - cq\}] = \exp(at)S(0)$, includes operating costs, depreciation, additions to accomodate growing demand, etc. To repeat, under the constant growth rate assumption we have $S(t) = S(0)\exp(at)$. Let the interest rate be the constant $r$.

### *I1: The social interest*

We take society's problem to be to choose t to

$$\text{maximize } PV(t) - F\exp(-rt), \tag{13.1}$$

where $PV(t)$ denotes the present value of the stream of net benefits from $t$ on. This quantity is computed by integrating the present value flow, $\exp(-rt)S(t)$ from $t$ to infinity. One obtains

$$PV(t) = \exp[(-r + a)t]S(0)/(r - a). \tag{13.2}$$

Note that $r > a$ for this to be finite. This makes sense. If net benefits grow faster than the rate of interest then the value of the project grows without bound. Differentiate (13.1) to obtain the first order condition for a maximum, and solution $T$,

$$S(0)\exp(aT) = rF, T = (1/a)\ln(rF/S(0)). \tag{13.3}$$

It is easy to check that $T$ is the only solution to the first order condition and it satisfies the second order necessary condition for an interior maximum. This completes the analysis of the deterministic case for the social interest. Note that larger $S(0)$ and smaller $rF$ gives earlier building time $T$. If $T$ is negative in (13.3) then the 'corner' solution, $T = 0$, is optimal. We shall assume interior solutions in the following.

### *I2: Monopoly*

For monopoly, profit, $P(t) = \exp(at)[\max\{D(q)q - cq\} = P(0)\exp(at)$. Copy the procedure above to obtain

$$P(0)\exp(aT) = rF, T = (1/a)\ln(rF/P(0)). \tag{13.4}$$

Since $P(0) < S(0)$ we have building at a later date than socially optimal. However, there is spillover at each date beyond the building date. The spillover not captured by the monopolist is the area $S(q^*) - D(q^*)q^*$ where $q^*$ is the monopoly level of output. This spillover grows at rate a after building the facility to produce for the market.

### *I3: Competition to capture a monopoly position*

Consider now a form of competition where there is a race and the first one to enter by paying $F$ at date $T$, captures a monopoly like the above from $T$ on. This forces

$$PV(T) = \exp[aT]P(0)/(r-a) = F, \quad T = (1/a)\ln((r-a)F/P(0)). \quad (13.5)$$

Here, as one would expect, building takes place earlier. The 'winning' monopolist builds earlier and captures no $PV$ profit for herself. She does this to avoid being displaced in the market by an earlier builder if there were still positive net $PV$ to be had. Spillovers of the amount $\exp(at)[S(q^*) - D(q^*)q^*]$ are captured by the public from the building date on.

### *I4: Competition to capture a perfectly discriminating monopoly position*

Consider a form of competition where the first one to enter by paying $F$ at date $T$ captures a perfect price discriminating monopoly where the operator can use nonlinear prices to sweep out the entire area under the demand curve and capture the entire net benefit at all dates after entry. This forces

$$PV(T) = \exp[aT]S(0)/(r-a) = F, \quad T = (1/a)\ln((r-a)F/S(0)). \quad (13.6)$$

Here we have an earlier date of building the facility than competing monopolists for the market in (13.5). Furthermore if $F$ measures the true cost to society as a whole of undergoing the cost of building, the entire capitalized net benefit from building is absorbed by the cost. This is truly a society that is enslaved by a rat race of building too early to service emerging demands. The discussion in Evans (1983, p. 219) shows how the welfare cost of this kind of rat race can be very large to the society as a whole. The 'dual' case of a rat race of adopting RCRs too early from a social welfare benchmark will be examined below.

### *I5: 'Actual' enterprise management in the 'real world'*

Given the above analysis of the social interest I1, the analysis of I5 is trivial. We simply replace (13.1) with (13.1′) maximize $(w1)PV(t) - (w2)F\exp(-rt)$, where $w1$ is the fraction of $PV(t)$ captured by the managers and $w2$ is the fraction of costs borne by the managers as determined by their incentive contracts. Much of the discussion on the New Zealand restructuring reforms reviewed by Lewis Evans *et al*. (1996) involved the design of appropriate ways of designing managerial contracts to provide an incentive to the managers to capture potential gains to the enterprise as a whole on the public interest standard rather than 'rent-seeking' the enterprise for themselves. Writers on this subject (e.g. Evans *et al*. (1996), Gibbons in Kreps and Wallis (1997) stress how the careful design of an accounting system upon which to write incentive contracts for management and workers is critical.

Equation ($1'$) is a crude attempt to capture this idea in a tractable way within the scope of this review.

Naturally if $w1$ is close to zero and the management has to bear the costs of unpleasant relationships with their workforce in gearing up to enter a new market or adopting different work practices to capture an RCR then obviously management is not going to feel motivated to do it. Indeed when one looks at the implied incentive contracts in some enterprises (not just the public ones) it is amazing that anything gets done at all. We encourage the reader to look at Lewis Evans *et al* (1996) for a thorough and vivid discussion of the New Zealand case.

To sum up, this deterministic analysis of building to service emerging markets, and bearing costs now to adopt RCRs which promise to deliver future gains, while trivial, does help us focus attention on cases where the rate of building may be too fast or too slow for the social interest, depending upon the industrial structure. This kind of analysis is easy to extend to include imperfect property rights, rule of law and theft, taxation, possible government confiscation without compensation, political instability, infrastructure (which impacts profit flow through cost of production), human capital (which reduces adoption cost $F$), subsidies to reduce $F$, etc. These factors impact adoption time and would feed into the forces emphasized by Barro (1997). Turn now to uncertainty in the growth of value produced by a facility once it is built.

The general theme will be that an increase in the instantaneous variance of uncertainty creates an 'option' whose 'option value' will lead to an increase in the expected present value of the problem and a shortening of the building time due to exercise of this 'option.' This theme has been explored by Brock and Rothschild (1986), Brock, Rothschild, Stiglitz (1989), hereafter 'BRS', Dixit and Pindyck (1994), Malliaris (1982), and others.

In order to get started, consider the problem

$$\text{Maximize } E\{\exp(-rT)(PV(T)-F)\}, \tag{13.7}$$

subject to the Ito stochastic differential equation,

$$\mathrm{d}S(t)/S(t) = a(S(t))\mathrm{d}t + b(S(t))\mathrm{d}Z, \quad S(0) = S_0 \text{ given.} \tag{13.8}$$

where $E\{\cdot\}$ denotes mathematical expectation conditioned on $S(0)$. Here one chooses the 'stopping time' $T$ as a function of past information to solve (13.7). To do this, one can first compute the quantity $PV(T)$ in (13.7), for each given $S(T)$ by using the Feynman–Kac formula as in Duffie (1988, p. 307). I.e. $E\{PV(T)\}$ such that (13.8) starting at initial condition $S(T)$ solves an ordinary second order differential equation with two boundary conditions. Doing this, one can write the quantity in the form $R(S(T))$. Then one may follow BRS (1989) and solve

$$\text{Maximize } E\{\exp(-rT)(R(S(T))-F)\}, \tag{13.9}$$

subject to (13.8), by using 'barrier strategies'. Do this by restricting oneself to the set of stopping times $T$ that stop the first time a barrier $S > S(0)$ is hit by (13.8) starting from $S(0)$. Hence the problem reduces to the scalar problem

$$\text{Maximize } L(S;S(0))[R(S)-F], \tag{13.10}$$

where $L(S;S(0))$ is the Laplace transform of first passage of (13.8) from $S(0)$ to $S$. The Laplace transform solves an ordinary differential equation of second order with two boundary conditions of particularly simple form. There is a subtle technical issue of locating sufficient conditions for the restricted class of barrier solutions to be optimal in the class of all stopping times adapted to the filtration generated by (13.8) but we refer the reader to BRS and references for that. Barrier solutions will be adequate for the illustration to be given here. Let us do the simplest case where $a(S)$ and $b(S)$ are constant.

Put $v$ equal to one half the square of $b$. The solution of (13.8) is lognornal, i.e. $\ln(S(t))$ is normally distributed with mean $m(t) = \ln(S(0)) + (a-v)t$ and variance $(2v)t$. It is convenient to transform variables by putting $X = \ln(S)$ in (13.8) to obtain

$$\mathrm{d}X = a'\mathrm{d}t + b\mathrm{d}Z, \quad a' = a - v, \quad X(0) = X_0, \tag{13.11}$$

which, as we said above, has solution $X(t)$ normally distributed with mean $X0 + a't$, variance $(2v)t$. Compute the expectation conditional at date $t$ on $X(t)$ of the integral from $t$ to infinity of $\exp(-rs + X(s))$ by the moment generating function formula for the normal $(E\exp(bZ(t)) = \exp[(2v)t])$ to obtain

$$R(S(t)) = S(t)/(r-a) - F. \tag{13.12}$$

Recall that we have changed units by putting $X = \ln(S)$ to get a linear stochastic differential equation for the dynamics of the state variable. Suppress $X$ in the notation for $L(X;X(0))$, and write $L(X;X(0)) = L(X(0))$. It is easy to show (See BRS (1989) or Malliaris (1982), for example) that $L$ satisfies the second order ordinary differential equation,

$$rL - a'L' - vL'' = 0 \tag{13.13}$$

The roots of the characteristic equation for (13.13) are both real with one positive and one negative. Since a Laplace transform is always positive and bounded above by unity only the positive root is sensible. Furthermore if $S(0)$ is already at the target barrier $S$, the Laplace transform is unity. Hence the solution of (13.13) is given by

$$L(X;X(0)) = \exp(P(X(0) - X), \tag{13.14}$$

where $P$ denotes the positive root of the characteristic equation for (13.13),

$$P = (1/2)\{-a'/v + [(a'/v)^2 + 4(r/v)]^{\frac{1}{2}}\}. \tag{13.15}$$

We may now write the value of a problem starting at $X(0)$ with barrier target $X$,

$$W(X(0);X) = \exp(P(X(0) - X))[\exp(X))/(r-a) - F]. \tag{13.16}$$

The procedure above has reduced a hard problem to a simple problem. For optimum problems one can now apply simple calculus to differentiate (13.16) with respect to $X$ and use the first and second order conditions for a maximum to locate optimal $X$. For entry competition problems, one finds $X$ that sets (13.16) to zero. One can now analyse the impact of parameter changes such as increases in variance '$v$' upon the entry times under the structures I1–I5. Here is a very brief sketch of how to do this.

First, assuming the economically sensible case $r/a > 1$, it is easy to show that $P(v)$ decreases from $P(0) = r/a$ to zero as $v$ increases from zero to infinity. Since $P(\cdot)$ acts like a discount rate, a decrease in the discount rate tends to make the problem worth more. An increase in variance '$v$' tends to make the optimal stopping size $X$ bigger. Increases in stopping size $X$ tend to make building occur slower, i.e. at later times, in order to capture the option value contained in the variance increase. Notice that one must restrict $P$ to be between unity and $r/a$ in order for the optimum problem to be well posed.

The mathematics sketched above now makes it possible to analyze the impact on problem values and X-sizes of parameter changes under different structures like I1–I5. We refer the reader to BRS (1989) and Dixit and Pindyck (1994) for details on how to solve these kinds of problems. BRS (1989) shows how to do local comparative statics such as local increases in variance for a wide class of these problems as well as do the mathematics of 'smooth pasting' and 'free boundary' problems in general. Malliaris (1982) has an especially straightforward treatment at an elementary level. Second, I1 gives a benchmark, the social optimum, against which the performance of other arrangements like I2–I5 may be measured. Turn now to the 'dual case' of RCR theory.

Let $F(t)$ be given by the Ito stochastic differential equation,

$$\mathrm{d}F/F = -c\mathrm{d}t + b\mathrm{d}Z, \tag{13.17}$$

and let the instantaneous standard deviation be zero on the growth of demand in the above problem in order to avoid dealing with multidimensional optimal stopping problems as in Brock and Rothschild (1986). Think now of $F$ as a cost to be paid which lowers unit cost of production by a fixed amount and this translates into an increase in profits (or social benefit) of $D$ to be computed below. Before going further we must make a digression on expectations, general equilibrium effects, and game theory.

We ignored general equilibrium effects and game theoretic modelling issues in setting up the five comparison 'institutions' I1–I5 on the 'primary side' treatment above. On the RCR side of the analysis we must confront head on the issue of how long the adoptor of an RCR expects to keep her new advantage before another competitor adopts an RCR and uses that advantage to undercut her.

We can only speculate about what a completely specified game theoretic 'common knowledge' model of adoption 'equilibria' would produce. One might, however, confidently predict that there will be 'surprises' like parallels to 'no trade' theorems for example. See Dekel and Gul's survey in Kreps and Wallis (1997) for a discussion of common knowledge frameworks in game theory and results like no-trade theorems. We wish to warn the reader about the complexities that a proper game theoretic treatment of adoption equilibria would raise. See Aghion and Howitt's (especially their references to patent race games and the like) piece in Kreps and Wallis (1997) for subtleties that should be dealt with, that we are ignoring here.

This complexity is simply avoided here by positing a random variable $K = T - t$ with density $Pr\{K = k\}$ that determines how long the RCR adoptor will be allowed to keep her advantage before she is displaced by someone else if she adopts an RCR at date $t$ and her advantage is snatched away at $T > t$. I.e. think of $K = T - t$ as the keeping time of the advantage. Let the expected value conditional on $t$ of this advantage be denoted by $g$ ($g$ for 'gain') which is assumed, for simplicity, to be independent of $t$. Then her problem is to choose a stopping time $t$ (adapted to the process (13.17) for $F(t)$) to solve,

$$\text{maximize } E\{\exp(-rt)[g - F(t)]\} \tag{13.18}$$

subject to (13.17) above. Notice that this problem would be analytically similar to that already treated above if '$g(t)$' was governed by the dynamics (13.8) and $F(t)$ were constant in $t$. I.e. $R(g(T))$ in (13.9) would be interpreted as the expected present value of incremental profits from the RCR which costs $F$ to implement. Thus the reader may use (13.8) and (13.9) as the solution procedure for an appropriate class of RCR timing problems as well as the market timing problems treated above. In view of this similarity we shall be very brief in the following exposition of an RCR problem that cannot be completely mapped into the treatment of (13.9) above.

Let $X(t) = \ln(F(t))$ and consider the Laplace transform of first passage from initial cost $X(0)$ down to $X < X(0)$,

$$L(X;X(0)) = \exp(P(X(0) - X)), \tag{13.19}$$

where $P$ is now the negative root of the equation (13.13) with $a' = -c - v$. As in the above analysis, we may reduce problem (13.18) to

$$\text{maximize } \{\exp(P(X(0) - X))[g - \exp(X)]\}. \tag{13.20}$$

Analysis of this problem is now a straightforward adaptation of the above analysis.

### *13.2.2 A summing up*

We have seen enough variations on the 'adoption problem' to sum up what we have learned. First, the structure I5 stresses the importance of designing an accounting system upon which award to bonuses to the management of the enterprise for the unpleasantness of bearing cost $F$ in order to get the gain '$g$' from an RCR or identifying and entering a new market. Much of the wave of privatizations and 'corporatizations' during the 1980s and 1990s was driven by the recognition that incentives to innovate were severely distorted by the regulatory framework in many countries. See, for example, Lewis Evans *et al.* (1996) for a review of the New Zealand 'experiments' in restructuring to induce efficiency enhancing incentives.

Second, the instantaneous variance in the rate of gain to new market entry or gains to RCR as well as the instantaneous variance in the rate of fall of adoption cost is a modelling way of capturing the presence of latent options the size of which increase with these instantaneous variances. The performance of the four different structures I2–I5 in capturing these opportunities (measured relative to the public interest standard I1) can vary widely. Perhaps the worst structure from the social welfare perspective is a variation of I5 where the management gains a trivial part of '$g$' but bears a nontrivial part of $F$ if they adopt.

Third, the list of parameters in the structures I1–I5 can be given more concreteness by looking at a particular industry such as agriculture where cost and benefit parameters to an individual farm of adopting a technology are influenced by the regulatory and tax framework; reference group adoptions; activities by government such as agricultural experiment stations and extension offices; relative input price shifts which induce innovations in the direction of economizing on relatively more expensive inputs as in the Ruttan theory of induced innovations; volatility in the ratio of benefits/costs to adoption of RCRs; and more. See Carlson *et al.* (1993) for an excellent review of agricultural and resource economics issues.

### *13.2.3 TFP growth, aggregate production functions, and growth regressions*

The reader may ask at this point: What does the analysis of RCRs and market seizing have to do with the dynamics of technical change that appear in the endogenous growth theories reviewed by Baumol (2001, this volume), Hahn (2001, this volume) and others at the Siena School, July, 1998? For example, let $X$ be a vector of inputs, let $F(X)$ be a constant returns production function, let $\mathrm{d}\log(A(t))/\mathrm{d}t$ measure TFP growth. Then given a set of input prices, assuming they are constant though time, suppressing them in the notation, we can write the unit cost function as $C(0)/A(t)$ so that the rate of RCR is measured by $|\mathrm{d}\log(C(0)/A(t))/\mathrm{d}t| = \mathrm{d}\log(A(t))/\mathrm{d}t$.

Consider the parameter changes and changes in industrial structure that make adoption times earlier in the analysis above. It is plausible to conjecture that the same changes are positively related to changes in at least the level of $A(t)$ (and perhaps the rate of growth, $\mathrm{d}A/A\mathrm{d}t$, over an appropriate interval of time). While the above analysis has to be aggregated up to the economy-wide level to be connected to the determinants of $A(\cdot)$ as a complete theory like Aghion and Howitt's work in Kreps and Wallis (1997) and their book (1998) has done, it is nevertheless suggestive of regressors that one might want to insert into growth regressions. The analysis is used to organize a list.

Think of the analysis of the time to build as not only the time to build to service more of an existing market, but also the time to build to service a latent or emerging market. One could think of the actual building to service a latent or emerging market as the production of a 'new' good. On the 'dual or cost' side think of the time to adopt an RCR that will increase profits by an amount *D* which grows at some rate after adoption. Consider the key parameters in the above analysis: (Primary or Revenue Side) (i) the rate of growth '*a*' of potential profits in an emerging market; (ii) the instantaneous standard deviation, '*v*' of those potential profits; (iii) the cost *F* of building a facility to service the emerging or latent market; (Dual or Cost Side) (iv) the rate of fall of adoption cost '*c*'; (v) the instantaneous standard deviation '*v*' of the rate of fall of adoption cost; (vi) the rate of growth of profit before and after adoption of the RCR.

The parameters and their effects help focus concentration on key determinants of growth on the revenue side and the cost side. First, the impact of infrastructure such as reliable transportation, communications, and power networks is hidden in the notation. Obviously it takes a modicum of infrastructure to even build a facility, much less use it to service a market and extract profit from this activity. Hence if this kind of public infrastructure is complementary with production and decreases in its quality and quantity take place, we can expect building to be delayed and RCR adoption to be delayed.

'Rules-of-the-game' embodied in good government and societal norms (stressed by 'social capital' in sociology (Coleman (1990)), Putnam (1993) in political science and, Alesina (1997) and North in Arthur, Durlauf and Lane (1997) in economics) are key to good growth performance. Crafts in Kreps and Wallis (1997) gives a broad ranging survey of research into the forces that drive disparities in growth performance.

Putnam's (1993) study of the regional governments of Italy comes close to a 'natural experiment' to study the differential impacts on economic performance of differential quality measures of regional governments. Putnam argues that cross regional differences in governments in Italy have dramatic impacts upon regional performance, not only in economic growth, but also in satisfaction of the residents with their lives. We shall list below a collection of things that governments can do that stifle growth. Doing the reverse would, of course, stimulate growth. To put it bluntly, government can be a macroparasite or a macrocatalyst when it comes to growth. For example if government is allowed to do the equivalent of randomly take the facility's profit income after it is built, this factor increases the 'effective' interest rate '*r*' and building is delayed. If government taxes profits at a constant rate, this factor multiplies the present value of after-tax profits, building is delayed and RCR adoption is delayed (if, as is realistic, the full cost, both material and mental, of RCR adoption is not off-set by tax deductions). If taxation of current profits is progressive, then the value of latent options on the revenue side may be lessened and building delays may result. If criminal activity by both citizens and government officials cuts profits or increases building costs then building will be delayed.

If government installs a set of rules and regulations that place burdens on RCR adoption by increasing the cost, *F*, then RCRs will be delayed. For example, government can impose rules that help workers and management resist adoption of RCRs and, hence, increase *F*. Government can impose burdens on foreign competition by closing off international trade via protectionism. But notice that while this closing off will reduce adoption time, it could also reduce adoption rat races so the social welfare calculation becomes subtle. By this point, the reader can, no doubt, list many ways, perhaps 1001 ways, that the incentives to adopt RCRs and to build facilities to enter and to service markets can be cut. Hence an economy may die (or grow) due to 1001 cuts (enhancements) in incentives. Putnam (1993) puts forth 12 measures of institutional performance of regional governments that enhance economic growth and satisfaction of its people. One can imagine building an index of quality of government (much in the spirit of Barro (1997)) following Putnam and inserting it as a regressor in growth regressions. Opening the mind even further towards finding the avenues of death of an economy from 1001 cuts in incentives shows up in Harberger's article (1998) by looking at more disaggregated levels than is usual in growth studies.

Harberger (1998) makes a strong plea to analyze the determinants of TFP growth (both positive and negative) at a much more disaggregated level than the national level, i.e. the industry level is better and the individual firm level is better yet. In our notation Harberger, imagines each individual firm continually solving little optimal adoption problems by choosing adoption times of RCRs in 1001 ways and each firm is embedded in an industry of firms doing the same thing.

### *13.2.4 Welfare, spillovers, adoption avalanches*

The five structures I1–I5 were introduced, not only to illustrate how the industrial structure impacts entry and adoption times but also to expose subtlety in welfare analysis. Notice that structure I4 illustrates a rat race scenario where building to service new markets and adoption of RCR's takes place too fast relative to the social interest. It should be emphasized that all of the analytics here are partial equilibrium analytics so one can not do a serious welfare analysis. Obviously the analysis should be carried out in a general equilibrium framework more like that in the surveys of Aghion and Howitt, Crafts, Jovanovic in Kreps and Wallis (1997) and their references to writers such as Grossman and Helpman, Lucas, Romer, Krugman and others working the general equilibrium approach, as well as Iwai (2001, this volume). A general equilibrium approach with heterogeneous agents along these lines that might be extended with optimal timing components under different industrial structures is that of Horvath (1998) who has an entry/exit component in his model for firms. One could imagine building a general equilibrium model by using building blocks like the asset pricing models studied by Altug and Labadie (1994) and Akdeniz and Dechert (1996) where there is a representative consumer who generates the stochastic discount factors which are taken as parametric to value maximizing firms and where the value maximizing firms are solving intertemporal allocation problems and optimal timing problems like those treated above. The representative consumer drives the asset trading market in claims to the earnings of these firms which, in turn, determines the stochastic discount factors taken as parametric to these firms. This structure is general equilibrium 'enough' to do welfare economics and general equilibrium analysis. I suspect that some of the insights gained from the optimal timing analysis above will still hold in this general equilibrium setting. Of course the simple analytics of the above treatment will no longer survive and quite possibly numerical methods as in Judd's book (1998) as used by Akdeniz and Dechert (1997) will have to be adapted to carry out this kind of analysis. But, at least the numerics are now within our reach.

Another future research project, more along the lines of complexity theory as described by the introduction to Arthur, Durlauf and Lane (1997) is to trace the linkages through the system of an adoption of an RCR by one firm. For example, it seems sensible that adoption of an RCR by one firm is going to put pressure on the other firms in the industry. If they fail to adopt the RCR, the adoptor can shave her price and capture all of the market (assuming perfect substitutes are being produced) up to her production capacity. If she builds additional plant up to market capacity at her shaved price, she could capture the whole market. The analysis above can be adapted to include pressure on others to adopt after the first adoptor uses a limit pricing strategy. Prescott (1998) stresses how relative price changes on competing goods can put pressure on firms within an industry to adopt RCR's rather like the pressure applied by a limit pricer with lower costs. Barro and Sala-i-Martin (1995) treat diffusion models of technical adoption as an impetus in TFP dynamics and discuss empirical issues raised in the estimation and testing of such models. More will be said about diffusion models later.

Complexity-based models of impulses propagated by locally interacting agents along the lines of Durlauf (1993) and Scheinkman and Woodford (1994) may be useful in thinking through the impact of an RCR adopted by one firm in a 'graph network' of imperfectly competitive firms where the strength of the graph link between firm $i$ and firm $j$ increases with the size of the cross elasticity of demand between $i$ and $j$. This linkage through a network of cross-elasticities is a factor determining the size of the 'avalanche' of RCR adoption across this network of firms set off by an RCR adopted by one firm who lowers her price. If enough imperfectly competing producers cut their prices, that could force a 'distant' firm ('distance' measured via the size of cross-elasticity of demand) to adopt an RCR earlier which, in turn puts pressure on its nearest competitor to adopt an RCR. One could then study the size distributions of such 'waves' of RCR's and see if power law scaling and thick tailed distributions appear as discussed by Scheinkman and Woodford (1994).

Here is another way to illustrate TFP dynamics in an instructive way. Harberger (1998) presents a construct which he calls a 'sunrise/sunset' diagram where he plots percentage of value added on the horizontal axis and percentage of total RCR on the vertical axis. He shows for four 5-year time spans, for US manufacturing, that the percentage of value added by the top ranked RCR achievers to achieve 100 per cent of the total RCR varies from around 12 per cent to almost 50 per cent and the industries that achieve this vary highly across the four 5-year time spans. He also shows that a lot of industries achieve negative TFP (i.e. their costs go up, not down) during his four 5-year periods. He reports on studies that show that similar patterns appear at the firm level as well as the industry level. Curiously this suggests a 'fractal' type of self similarity structure in TFP dynamics. That is to say that the relationship of TFP at the firm level to TFP at the industry level is rather like the relationship of TFP at the industry level to TFP at the level of the whole economy. This 'self' similarity of structure of sunrise/sunset diagrams at different levels of aggregation is useful discipline for theory. For example, Harberger states that the 'mushroom' pattern revealed in his sunrise/sunset diagrams suggests that the kind of external spillover (if any) in propagating TFP dynamics is much more likely to be a localised spillover rather than a global spillover. This insight already sheds light and discipline on how one should enter spillovers in theoretical growth models in order to generate simulated sunrise/sunset diagrams consistent with Harberger's evidence. Returning to the issue of useful tools to detect evidence of spillovers, one might use Harberger's construct to see if there is any difference in the TFP dynamical pattern found as revealed in sunrise/sunset diagrams across different historical periods, in particular, the current period, which many commentators think is 'different' due to the huge potential spillovers from the nonrivalry and nonexcludability of the 'information goods' being produced in todays leading sectors. Recall that Baumol (2001, this volume) stresses the apparently large size of spillovers of benefits from inventors and creators to the population of past technological revolutions as well as the present, even though the earlier innovators did not have to deal with problems of nonrivalry and nonexcludability as much as, perhaps, in current times.

For example, DeLong (1998) reviews, from a long term historical perspective, whether there is any difference between today's 'leading sectors' of growth such as the computer industry centered in Silicon Valley and the biotech industries centered near university complexes. He argues that there is not an order-of-magnitude difference in TFP growth between these current media-hyped leading sector industries and leading sector industries of the past when TFP is properly measured as the number of hours a median worker must work to purchase a unit of service flow. The prosaic service of casting light into a dark room is a useful example to gain perspective. Think of the dramatic drop in number of median worker hours needed to purchase a unit of this service during the 'Edison' period. However, DeLong (1998) does say that there is a major difference in the degree of (i) rivalry, (ii) excludability, and (iii) transparency in the goods being produced by today's leading sectors in contrast to leading sectors in the past.

This difference in degree of three essential components of private property, stressed by DeLong, may translate into an increase in the degree of spillovers in the adoption models sketched above. This concern with spillovers due to an increase in the permeability of property rights brings us to the use of random field models, such as Ising models (Durlauf (1993)) in cycle and growth modeling and related empirical work using such models. Brock (1993) shows how useful approximations to the complicated probability structures of random field models, called 'mean field approximations,' may be introduced in such a way that one can borrow from the large literature in discrete choice econometrics in order to do empirical work with interacting systems models. See also Durlauf's and Ioannides's papers in SFI (II) for more on this. A central message of random fields models and their mean field approximations is this: as the degree of complementarity increases in these frameworks, alternative stable states appear and nonergodic behavior appears at a critical level of the degree of complementarity.

Hence, we can imagine building a model where there is a 'lattice' of 'sites' at which are located economic entities facing adoption decisions as in Durlauf (1993) or there is a graph structure as in some of the papers in SFI (II) that describes the network of potential spillovers amongst adoptors. I.e. let there be an arrow from agent *i* to agent *j* if *i*'s adoption lowers the cost to *j*'s adoption and let the 'size' of this arrow denote the strength of the cost reduction spilled over from *i* to *j*. Let each adoptor solve optimal adoption timing problems like the above but ignore her effects on the others connected to her. Brock (1993) and Durlauf (1993) discuss mathematics of these kinds of models in the simpler setting where the decision is binary: Adopt or Do not Adopt. The treatment of optimal timing to adoption in such interacting settings will surely be complicated.

However, one could imagine building general models that contain modules from the received innovations literature such as Aghion and Howitt, Jovanovic in Kreps and Wallis (1997), Brock's (1993) work using Curie–Weiss models and mean field approximations to general interacting particle systems models, Durlauf's (1993) work using random fields models that might possibly isolate a critical degree of spillovers due to changes in the degrees of rivalry and excludability in leading sector industries as they impact the rest of the economy. This kind of modeling might shed light on the consequences for general TFP performance for the economy as a whole of changes in the degree of rivalry and excludability in particular industries, especially leading sector industries. At the very minimum just thinking about carrying out such a modeling exercise motivates work on the econometrics of spillovers to be discussed below.

## 13.3 Complexity theory: any help?

In Section 13.2 we set out a context for evaluation of what complexity theory has to offer the study of cycles, growth, and structural change. We organized the discussion in terms of optimal timing for adoption of an RCR and briefly mentioned the importance of thinking in terms of a hierarchy of time scales and 'spatial' scales from the slow to the fast in time and from the large to the small in 'space'. The study of dynamics on smaller scales tends to require more disaggregated data.

Recall that Harberger (1998) stressed a disaggregated approach to understanding dynamics of TFPs and RCRs and used the expository device of 'Yeast versus Mushrooms' to capture the idea that TFP/RCR progress in any given decade is concentrated in a handful (mushrooms) of industries rather than being spread out more evenly across the whole economy. At the risk of repeating what was said in Section 13.2, one could imagine constructing Stommel type diagrams where the 'spatial axis' is not 'geographic space' but 'space' measured as an index of similarity across industries. A spectral version of a Stommel diagram does not strike me as useful in this context as a version where the temporal scale measured the size of the TFP movement at the industry index, where the index would be measured by something like an SIC code but quantised to stress product similarity. The pictoral scaling of TFP movement on Stommel plots may help locate patterns that would not be easy to see otherwise. See Clark (1985) and his references for the very real technical problems of constructing Stommel diagrams and see Gunderson *et al*. (1995) for some samples of the helpfulness of Stommel diagrams in pattern detection in applications. Although we have been purposely vague in exact details describing how to construct a Stommel diagram one can imagine creating such constructs after looking at the examples in Gunderson *et al*. (1995) and using them on 'spatial/temporal' ('space' being interpreted broadly) data sets to test the hypothesis as to whether the data is generated by a 'sandpile' type threshold-cellular automata type model used by Scheinkman and Woodford (1994), for example. If one plotted the time to relaxation and the 'size' of 'avalanches' generated by such models on a Stommel diagram, one might see 'blobs' at all size scales on the diagram.

In contrast 'Rnet' models stress two or three critical scales of time at which the bulk of activity 'clumps' with corresponding 'clumping' on spatial scales. See Gunderson *et al*. (1995) for explanation of this reasoning, case studies where it appears to fit, and Stommel diagrams that illustrate it. Notice that 'Rnet' type models are essentially stochastic differential equation models whose underlying deterministic differential equation (called the 'skeleton' of the model) with a two or three level hierarchy of time scales where the slower time scale movements serve as bifurcation parameters for the faster scale below it in this hierarchy. 'Spatial' dynamics are linked by dynamical spillovers in these models. We have said enough about the 'theory' of pattern formation. Turn now to the 'econometrics' of pattern recognition.

## 13.4 Pattern recognition

In the first SFI Volume on the Economy, SFI (I), there was some attention focused on the problem of pattern recognition including the problem of testing for patterns in noisy data, especially left-out structure in forecast errors of fitted models. Let me give a quick sketch of this development up to the present following Brock's papers in SFI (I), SFI (II) and add a bit on newer work.

Suppose a model is fitted to data and the forecast errors $\{e(t)\}$ are saved for testing. For example, consider growth regressions as reported in Barro (1997) and Durlauf and Quah (1999). Under the null hypothesis that the model being fitted is the true model up to Independently and Identically Distributed (IID) errors, then one can use a test for independence on these errors to specification-test the null model. However problems arise because the true errors are not known, only the estimated errors are available for test.

Brock, Dechert, Scheinkman, and LeBaron (1996), hereafter, BDSL, developed a test for IID errors which has the same first order asympotic distribution on estimated errors as on the true errors (under regularity conditions typically satisfied in applications). Their theorem takes care of the problem, that only fitted model errors are available to test, but there is another problem. In most applications forecast errors of fitted models (and perhaps the 'true' models) tend to display persistence in variance (called autoregressive heteroscedasticity). So applicators of BDSL's work tend to estimate a model of this error variance persistence and 'standardize' the errors so that if the model is 'true' these estimated standardized errors should be approximately IID.

Here is an example directly relevant to cycles and growth issues being discussed here. Altug *et al*. (1999) tested for linearity of detrended real per capita US GNP, call this data $\{x(t)\}$, and adduced evidence that $\{x(t)\}$ was not generated by a linear stochastic process. I.e. $\{x(t)\}$ is not a sample from a stochastic process that has a Wold representation with IID 'drivers' ('innovations' in technical language). They used several tests for nonlinearity including BDSL (1996). The pattern of rejections of linearity suggested the rejection was not just due to persistent volatility of the fitted model errors.

They computed 'Solow Residuals' (a conventional measure of TFP), call these $\{s(t)\}$, tested the hypothesis that $\{s(t)\}$ came from a linear stochastic process, and failed to reject. This lead them to examine the aggregate labor market where they adduced evidence that the dynamics appeared to be asymmetric. This lead them into an exploration of 'labor hoarding' type models, where there is an asymmetry in cost between hiring and firing a unit of labor input, a review of the literature on such models, and simulation of such models to see if such models could produce simulated data consistent with their findings. While the jury is still out on whether linear TFP dynamics coupled with asymmetric dynamics in aggregate employment dynamics is consistent with the aggregate data used in their study, their study illustrates potential usefulness of the kind of tools we are discussing here. See Granger and Terasvirta (1993), Pesaran and Potter (1992), and Potter (1995) for methods in nonlinear time series econometrics that are very useful for the study of cycles and growth.

See Dechert (1997) for other examples of applications of nonlinearity tests, including some applications to cycle and growth empirics. For practical guides to nonlinearity testing see Dechert's (1998) and LeBaron's (1998) websites for details, references, practical hints for applications, software discussions, research findings, etc.

Testing procedures like the above tests for 'left out structure' of fitted model residuals generated much evidence for 'departures' from 'conventional' models but did not give good guidance towards the causes of these 'departures.' So this evidence generated a flurry of activity towards designing sophisticated procedures which used computational inferential procedures such as variations on bootstrapping to produce null model distributions of statistics generated by the purposive economic behavior being modeled, such as profits from trading strategies and means and variances from returns to trading strategies.

The bootstrap could be used to bypass analytical impossibilities of calculating null model distributions for such complicated econmically motivated statistical objects such as trading profits, that are much more germane to the purposive economic behavior that the null model purports to be describing. This procedure has been pursued by Brock, Lakonishok, and LeBaron, (1992). See LeBaron's website (1998) for much more on this. The bottom line is that many conventional models overpredicted volatility of returns and under predicted returns following buy signals. If one tested the Efficient Markets Hypothesis (EMH) by testing whether past returns help predict future returns by averaging the mean squared error of prediction over all periods, one would miss these 'pockets of predictability'. However, it is not clear that this kind of evidence is a rejection of the EMH or is simply evidence of mismeasured risk. Indeed it is not even clear that this evidence (based on 100 years of daily data up to 1986) holds up on post-1986 data (cf. Sullivan *et al*. (1999)). See Acar and Satchell (1998) for papers carrying out related work (including one by LeBaron).

Authors doing this kind of pattern recognition work were concerned about 'data snooping' biases. Data snooping biases arise from researchers and the research community at large experimenting with different candidate models or trading rules over the huge space of potential models or trading rules. To put it another way, the research community, sharing a data base, can not avoid pre-test bias that comes from fitting different models using that data base and spreading this knowledge to a particular researcher who does a 'specification search' and selects a particular model to be fitted to that data. The point is that the conventional test statistics used to assess the quality of 'fit' of that model have contaminated null distributions because of the communal and individual specification searches.

A major way to deal with this problem is out-of-sample testing provided that the data used for the 'out-of-sample' testing is 'truly' out-of-sample and is not just a 'hold-out' sample saved for 'out-of-sample' testing. Even hold-out samples may have been inadvertantly snooped during the model selection phase. Recently work by Sullivan, Timmerman, and White (1999) has presented methods that build on recent advances in statistics that deal with this data snooping problem. It would be desirable to apply the set of tools reviewed above to empirical work in cycles and growth, especially the work on testing for 'left out structure' of received growth regressions. Even more important would be to adapt the work of Sullivan *et al.* (1999) on correcting for data snooping biases in received growth regressions. However, there is a serious problem in using such data hungry procedures in a field where data is scarce. Financial markets generate data at the 'tic by tic' frequency but cycle and growth empiricists must make do with quarterly and yearly data with a few advanced countries producing monthly data. Nevertheless, even a modest attempt to correct for data snooping biases using the above tools on growth regression exercises like those reviewed by Durlauf and Quah (1999) may produce useful information on reliability of the findings of that literature.

Indeed Durlauf and Quah (1999) are quite critical of received work on growth regressions partly because of the rather large number of regressors (relative to the size of the underlying data set) that have appeared in growth studies. One could do, in principle, a simulation exercise inspired by Sullivan *et al.* (1999) to correct 'significance tests' for 'regression fishing' of the type that concerns Durlauf and Quah (1999).

Another part of their critique concerns inadequate attention being paid by students of growth in addressing plausible alternative hypotheses such as threshold nonlinear models where the poor do not eventually catch up with the rich in contrast to conditional convergence models where they do eventually catch up (after conditioning on the same regressor set in both types of models). Policy prescriptions differ drastically across these two sets of models. Indeed much in the spirit of 'lump analysis' in ecology (cf. Allen *et al*. (1998)), Durlauf and Quah apply 'lump detection' methods from statistics to adduce evidence of multimodality (cf. their Figures 1, 7, 11a,b and surrounding text) to adduce evidence in support of lack of convergence even after conditioning on conventional regressors that indicate conditional convergence in 'linear' studies. It would be interesting to compare the lump detection methods of Allen *et al*. (1998) in ecology with those of Durlauf and Quah (1999) to detect evidence of possible multimodality in economic data, not only in the raw data, but especially after conditioning on received regressors that indicate 'conditional convergence clubs'.

## 13.5 Complex econometrics

We have given the reader a foreshadowing of the econometric issues raised by the use of complexity-based methods in economics and finance. Brock (1993) and my paper in SFI (II) gives a discussion that relates tools such as statistical mechanics, interacting particle systems theories, mean field theory, non-ergodic models, and ensemble analysis to received discrete choice econometric methods. Pure theory of social interactions and the complementary econometric theory is reviewed Brock and Durlauf (1995), (1999). We give some speculations here on possible use of these tools for growth empirics. The 'new' growth theory makes much out of spillovers whether due to production and technology as in Romer's work or to human capital as in Lucas's work (cf. discussions in Barro and Sala-i-Martin (1995), Durlauf and Quah (1999), Hahn (2001, this volume), Hall (1996)). These kinds of spillovers have a 'social multiplier' in the sense that benefits to a spillover originator multiply out across to others in a type of chain reaction or percolation throughout the whole system. The potential presence of such spillovers raises the econometric issue of identifying such spillovers and separating them from other effects that are essentially observationally equivalent to spillovers but have no social multiplier. See Manski's paper in SFI(II) for a review of his pioneering work on this problem.

In view of the central role of spillovers and the resulting lack of Pareto Optimality in the 'new' growth theory, it is surprising to find so little written on the practical econometric problem of identifying and measuring such spillovers using econometric methods. See Barro and Sala-i-Martin (1995, Chapter 8, especially their discussion of how spillover effects can lead to a dependency of country *i*'s growth rate on country *j*'s variables, p. 275) for work on estimating diffusion type models. Baumol's work (2001, this volume) makes such a good case for the existence and size of such spillovers that, perhaps, formal econometric work is not necessary to make a persuasive case for their existence, but it will still be useful for measuring their size and importance.

In any event we sketch how tools reviewed in Brock and Durlauf (1999) might be used to help identify and measure spillover effects in growth regressions. Consider Barro and Sala-i-Martin's (1995) and Barro's (1997) growth regressions and the framework laid out in DQ (1999) especially (DQ.19) in Section 13.2.

Following Manski in SFI (II) we say an 'endogenous' social interaction or 'true spillover' appears in a set of growth regressions in a cross section of economies (e.g. US counties (Wheeler (1998)), Japanese prefectures, regions in Europe, US states, and countries (Barro and Sala-i-Martin (1995), regions of Italy (Putnam (1993)) if an average over some set of economic entities, call it $S$, of an outcome variable appears in the right-hand-side of some of the regressions. Hence, if we averaged the left-hand-side of the outcome equations of these economic entities in $S$ over $S$ we have a set of simultaneous equations to solve for the outcome average over $S$.

This is a type of simultaneous equations problem which can lead to a lack of identification of parameters when estimation is done. Manski's paper in SFI(II) stresses that, in other applications, this problem raises fundamental issues of econometric identification of 'true endogenous' effects (those which have a social multiplier potentially exploitable for the public good by appropriate policy intervention) and 'spurious' effects (those which have no such multiplier and where beneficial policy intervention that exploits autocatalytic effects and positive feedback loops is not possible). Examples of 'true endogenous' effects include production spillovers via learning-by-doing, human capital spillovers due to average economy-wide human capital appearing as an argument in each firm's production function, and learning spillovers from highly skilled to new learners as in Glaeser's (1997a,b) city models.

Examples of 'spurious' effects that might generate false evidence of true endogenous effects if omitted in regressions include unobservables that are correlated across an economy that increase productivity to individual firms in that economy. Furthermore productive variables at the individual firm level whose economy wide averages also improve productivity at the firm level, but yet have no social multiplier also can cause identification problems as pointed out by Manski.

Here is the simplest example I can think of to expose the problem quickly. Following Manski's paper in SFI (II), suppose $y$ is the outcome variable of an economic entity which is of interest, $E(y|g)$ is the average of $y$ in reference group $g$, $x$ is an input variable that effects $y$, $E(x|g)$ is the average of $x$ in reference group $g$ which may also impact $y$. Consider the regression in levels, where $E(e|x,g) = 0$,

$$y = a + bE(y|g) + cx + dE(x|g) + e. \qquad (13.21)$$

To obtain a social 'rational expectations' equilibrium reduced form, take the average of both sides with respect to reference group $g$ in (13.21) to find $E(y|g)$ and insert that back into (13.21) to obtain the reduced form

$$y = A + cx + DE(x|g) + e, \qquad (13.22)$$

where $A = a/(1-b), D = (bc+d)/(1-b)$

Notice that a social multiplier, $[1/(1-b)-1] = b/(1-b)$, exists for the regressor $E(y|g)$. I.e. if, for example, the constant term '$a$' increases by an amount $f$ for one member of a group $g$ of size $|g|$, the group mean, $E(y|g)$ increases by a factor $f/[(1-b)|g|]$ for a net gain of $bf/[(1-b)|g|]$ over the case $b=0$ of no multiplier. This is a measure of the size of spillover. We see the source of the econometric problem raised by (13.21) and (13.22) right away. There are only three regressors in the reduced form (13.22) but there are four parameters to estimate in the primary form (13.21). The coefficient of interest, '$b$', cannot be separated from $d$. Taking first differences does not solve the problem because the constant is lost. We still have one more parameter relative to regressors in the reduced form.

It is easy to see how this kind of situation can arise in a growth regression exercise inspired by the spillover models of Romer (spillovers to our economic entity's output, $y$, across outputs in reference group $g$ due to learning by doing spillovers) and Lucas (spillovers from average human capital in reference group $g$) (discussed in Barro and Sala-i-Martin (1995), Barro (1997), and Durlauf and Quah (1999)).

Following Brock and Durlauf (1999), we consider four avenues of potential escape from this problem. First, one might be able to use economic theory to argue that a term like $cx+dE(x|g)$ should not appear in the regression equation and that it should be replaced by a term like $cx+dE(w|g)$ where $w$ may be correlated with $x$ but is not identical to $x$. This escape route leads to a reduced form with the same number of regressors as parameters but the '$X'X$' matrix of the regression may be near singular if the correlation between $x$, and $w$ is strong enough.

A second avenue is to attempt to use economic theory to argue that the linear combination of $x$, $E(x|g)$, should be replaced by, for example, a linear combination of nonlinear functions of $x$, $E(x|g)$. In this case the '$X'X$' matrix would be nonsingular except for 'hairline' cases. However the $X'X$ matrix may be near singular if the nonlinearity was weak or the support of $x, E(x|g)$ was narrow so that there was not enough variation to let the nonlinearity speak loudly enough to the nonsingularity of $X'X$. A third route of escape is to replace the current values of $E(y|g), E(x|g)$ in the regression (13.21) above and replace them by their lagged values or, perhaps, some function of their lagged values. See Manski's paper in SFI (II) for a critique of this route of escape as well as some other routes in other applications than cycles and growth applications. His critique may, however, be of concern to researchers in cycles and growth. A fourth route of escape is to consider other ways of adducing evidence for spillovers than growth regressions. For example if one were studying TFP dynamics at the individual firm or individual industry level, then one could define a relevant comparison group $g$ and attempt to estimate spillovers from the average adoption time of a particular innovation of the reference group $g$ upon the time of adoption of the individual firm. Turn now to one possible way to implement this strategy.

### *13.5.1 Hazard model econometrics and measuring spillovers in adoption times*

A fundamental concern in measuring whether there are spillovers across economic units in TFP dynamics is the problem of estimating the impact of spillovers (e.g.

the impact of the average time to adoption in a relevant 'peer group' sector to individual firms in that sector). We sketched above how optimal timing theory suggests forces that shorten the time to adoption of an RCR to an individual adoptor. If there are 'true spillovers', for example a decrease in the average time to adoption of a 'reference group' of firms decreases the cost $F$ to adoption for an individual member of that reference group, then we expect from the theory sketched above that our firm will adopt earlier.

A social adoption equilibrium can be defined in terms of hazard functions as follows. Let $T$ be a non negative random variable giving the adoption time and let $T$ have absolutely continuous cumulative distribution function $F(t) = Pr\{T < t\}$ and density function $f(t) = \mathrm{d}F/\,\mathrm{d}t$. Then the hazard function, $h(t)$, is defined by

$$h(t) = f(t)/[1 - F(t)]. \tag{13.23}$$

Since the survivor function $Pr\{T > t\}$ is the exponential of the negative of the integral of the hazard up to $t$, therefore once the hazard function is specified, we may write $F(t)$ and $f(t)$ in terms of the hazard. Consider an example of individual firms in an industry potentially adopting a piece of cost-reduction technology. Imagine one has a data set on firms and industries that contains individual characteristics, $x$, of the firms, industry characteristics, $w$, (some of which are just averages of firm characteristics within each industry), and adoption times for different cost-reducing practices. Suppose one wishes to adduce evidence for or against existence of 'true spillovers' of adoption times of each firm within an industry from the rest of the firms in that industry using this data set. Here is a possible strategy. Write down a hazard function $h(t,x,w;E(t|g))$ for each individual firm as a function of individual characteric vector $x$, its industry characteristic vector $w$, and the firm's expectation $E(t|g)$ on the average adoption time of firms in that industry. Here '$g$' denotes the industry of the firm. The actual average time to adoption in that industry is the average over the set of firms in that industry of the individual firm adoption times (assume all firms eventually adopt). If the term, $E(t|g)$ did not appear in the individual firm's hazard, we could simply apply standard methods of hazard function estimation (cf. Heckman and Singer (1985)) and estimate the hazard function. We illustrate how to modify the standard methods to include social interactions in the hazard.

Consider the simple example of exponential hazard constant in $t$. Then Et is just the reciprocal of the hazard. Now specify a functional form, $h(x,w;E(t|g))$ for this hazard in 'regressors' x,w,E(t|g). We have by the properties of exponential hazards

$$E(t|x,w;E(t|g)) = 1/h(x,w;E(t|g)) \tag{13.24}$$

Impose 'rational expectations' for each reference group $g$,

$$E(t|g) = \text{ sum } \{1/h(x,w;E(t|g))\} \text{ over members of } g. \tag{13.25}$$

Equation (13.25) defines a rational expectations equilibrium. Brock and Durlauf (1999) show that if the hazard is known up to a finite dimensional parameter vector, then maximum likelihood may be used to estimate (13.24) subject to (13.25). For a simple example consider,

$$h(x, w; E(t|g)) = \exp[-(a + bx + cw + dE(t|g))]. \tag{13.26}$$

One may locate sufficient conditions for existence of a rational expectations equilibrium (i.e. a solution to (13.25)). When one checks the identification condition for $(a, b, c, d)$ to be identified in (13.26) one ends up looking at the matrix of second order partial derivatives of the likelihood with respect to the parameter vector. This gives one a matrix that looks rather like a regression matrix $X'X$. A sufficient condition for identification requires that this matrix be nonsingular subject to solutions of (25).

If one has a data set on individual firms and industries where there is variation in adoption times across individual firms and there is variation in mean adoption times across industries then the identification condition will be typically be satisfied even if 'regressor pairs' like $(x, E(x|g))$ appear in the hazard. Notice that regressor pairs $(x, E(x|g))$ destroy identification of 'true' spillovers in the linear case (13.21), (13.22). The reason for this happy identification result in the hazard case is the nonlinearity of the hazard and the rational expectations equilibrium condition keeps the identification matrix from becoming singular on the set of solutions to (13.25), unlike the linear case (13.21), (13.22). See Brock and Durlauf (1999) for the details. We wish to warn the reader at the outset that carrying out this strategy of adducing evidence for 'true spillovers' in adopting RCRs may be difficult to carry out in practice. The main point we wish to make here is this. The application of optimal timing theory above and thinking about interactive relationships amongst firms solving optimal timing problems suggests pathways for 'true spillovers' to operate. This, in turn suggests econometric exercises using estimation techniques from duration analysis which are discussed in Heckman and Singer (1985). Our sketch of extending these techniques to interactions models is intended as an illustration of what future research using these tools might look like. Turn now to another area of potential application of tools discussed here.

### *13.5.2 Expectations: rational or other?*

The debate on whether it is more useful in cycles and growth research to use rational expectations modeling or some type of adaptive or boundedly rational modeling was quite intense during the Siena School. See Amendola and Gaffard (1998), Hahn and Solow (1995), and Flaschel (2001, this volume), for alternatives to rational expectations modeling. In this section of the chapter, I discuss whether some recent work might contribute to a narrowing of the disagreement among scholars. There are several potentially useful approaches to use data and 'complexity-based evolutive' type theory to help narrow disagreements in this area. First, consider the use of asset pricing data in the countries that have well-developed asset markets. One can insert an asset pricing module into their favorite macro model that departs from rational expectations and crank out the implied time path of stock prices and trading volume and ask whether these look like aggregative stock price data and aggregate trading volume data. This strategy allows one to check whether their favorite macroeconomic model generates implied asset prices consistent with empirical findings from the large event studies literature in finance. This strategy imposes useful discipline on model building. Some examples are given in my paper in SFI (II). Here are some more.

For example, rational expectations type asset pricing models like those with a production side with tax distortions reviewed in Altug and Labadie's recent text (1994) generate the following testable (in principle) prediction. If one has a sufficient state space dynamical representation for the dynamics of the 'real side' of the model then here is a sharp empirical implication. Past asset prices do not help incrementally predict future asset prices given past and current values of the state variables of the 'real side'. Hence one can do a nonparametric regression of future prices upon past prices and past state variables and set up a test of 'significance' of past prices in this regression given past state variables. If one fails to reject the null hypothesis that past prices do not enter this regression, then one has adduced evidence consistent with the hypothesis of rational expectations in this context. This procedure is sometimes called 'Granger causality testing' in the time series literature. Parenthetically, we wish to emphasize that this strategy is not nearly as easy to carry out in practice as we are making it sound in this brief exposition. We are burying a lot of problems in order to expose the basic structure of the strategy. One major problem is unmeasured components of the state space representation that impact future earnings which the stock traders see but the scientist does not. These unmeasured components could be correlated with the measured components and this correlation induces a pathway through which past prices might help predict future prices. Hence a researcher might erroneously reject the rational expectations hypothesis when it is true. In any event let us continue with our discussion.

Second, we may be able to adapt a strategy used in the debate on the causes of mean reversion (if any) in asset prices. Jog and Schaller (1994) argued that mean reversion is likely to be due to liquidity constraints and differential access to raising funds in the financial markets that especially bite on the smaller firms relative to the larger firms. This hypothesis had distinct implications in a production based rational expectations asset pricing model where some firms were liquidity constrained and others were not. For example this kind of economic force would lead one to expect mean reversion to loom larger in smaller firms and to loom larger during periods of economy wide liquidity stress such as deep recessions. Indeed the evidence for mean reversion appears weak for large firms post-World War II in contrast to small firms during deep recessions. See Jog and Schaller (1994) for the details.

Third, one might 'back-off' from rational expectations in the direction of evolutionary dynamical models where ideas from evolutionary computation theory are used to set up models where expectational schemes co-evolve while competing with each other in a 'Digital Darwinistic' struggle for existence. See Arthur *et al.*, and Darley and Kauffman in SFI (II), LeBaron (2000), and Sargent (1993) for work of this type. See, especially LeBaron's review (2000) for an excellent picture of this area. One can contrast the prediction of no 'Granger causality' of past prices for future prices in rational expectations models above with simulated data from evolutionary computational models to see if one gets a Granger causal pattern that mimics that found in actual asset market data.

Fourth, one can estimate a class of models containing a mixture of rational expectations agents and boundedly rational agents as did Baak (1999) for the case of cattle ranchers and did Chavas (1999) for the case of pork producers, where the data are free to speak to the presence of boundedly rational agents. Baak, for example, was able to test, using data on the cattle industry, whether the extra 'free parameters' brought by two different types of bounded rationality, naive simple backwards looking predictors, and Nerlove's Quasi-Rational predictors were 'significant.' His evidence suggests some support for Nerlove's level of rationality in contrast to fully rational expectations.

Nerlove's quasi-rational expectations impose some level of rationality for instance being consistent with historical features of the data such as the autocorrelation structure of past prices. This idea is related to Kurz's Rational Beliefs (1997) which impose rationality to the level of consistency with the data but do not require agents to possess full knowledge of the structure of the world that they live in. Kurz adduces evidence to support his type of departures from full rational expectations. Baak's evidence suggests that boundedly rational ranchers left a detectable trace in the data. However, his rejection of the null hypothesis of purely rational expectations ranchers may have been due to specification of a false rational expectations model on a truly rational expectations sector even though he used the well-known Rosen, Murphy, Scheinkman model for his null model. His rejection could also be due to unintended data snooping bias of the type discussed by Sullivan *et al.* (1999). But, it is encouraging that Chavas (1999), using different methods than Baak, obtained rather similar results using data on the pork industry.

Fifth, Brock and Hommes (1997), (1998) develop a theory of 'Adaptively Rational Equilibrium Dynamics' (ARED) where economic agents hold a set of different belief systems and models of the economy and use a discrete choice model to select their own belief out of this set based upon performance indices. The performance index for a particular belief system or predictive model is built out of past profits gained from using that system, for example, a distributed lag of past profits. Rational expectations are also available (for a fee). Costs of obtaining information to implement a belief system (e.g. rational expectations) are subtracted from the profits gained by using it. The system of beliefs co-evolve in ARED theory in an evolutionarily dynamic equilibrium.

The ARED theory is rather like the computationally-based work of Arthur *et al.*, Darley and Kauffman in SFI (II), and LeBaron (2000), but is framed in a setting where analytical results can be obtained and econometric work can be done. In this way the massive literature on discrete choice econometrics and rational expectations econometrics can be adapted to do econometric work in testing for the presence of expectational schemes being used by economic agents other than rational expectations. The work of Baak and Chavas discussed above is an example of recent econometric work that allows data to speak to the presence of boundedly rational agents.

In ARED theory one introduces evolutionary competition of different belief systems in generating trading profits net of costs of obtaining the information needed to implement those belief systems. Rational expectations will be acquired by economic agents only when there appears (based upon the recent past) to be enough value to the extra predictive accuracy to cover the cost of obtaining rational expectations. So the theory 'backs off' from fully rational expectations but does not abandon the intelligence of agents. Hence there will be 'phases' where cheaper boundedly rational expectations are used by the bulk of agents. In these 'phases' past prices (in most formulations of such mixed expectational models) will incrementally help predict future prices given past and current values of the state variables of the 'real side.' I.e. past prices will Granger cause future prices during these temporal phases in this setting. This implication may allow one to use Granger causality testing methods in time series econometrics to test for the presence of boundedly rational traders.

Whether any evidence of departure from rational expectations is detected by empirical testing will, of course, depend upon the size of the phases where rational expectations are abandoned by the bulk of agents because they are not worth the cost of obtaining. Notice that if empirical evidence of departures from rational expectations is easy for econometricians to find in this kind of model the very logic of the model says that rational expectations are likely to be used by most of the agents (unless, for some reason, the costs of obtaining rational expectations are very large). In turn the size of such phases will depend upon the tradeoff between the gains to the extra 'look ahead' value of rational expectations in improving profits to the costs of obtaining such sophisticated expectations.

During stable periods where the state of the system is not changing very much, one might expect the gains to exploiting rational expectations not to cover their cost in which case the agents revert to naive schemes. Nevertheless, even a modest amount of predictability out-of-sample of future returns using past returns is inconsistent with a lot of empirical evidence in asset markets, especially heavily traded assets in well developed markets. Yet, common knowledge rational expectations frameworks tend to lead to small trading volume whereas trading volume in reality is very large. This lead to development of a related type of modeling strategy that is still evolutive. See LeBaron (2000) for an excellent review of the latest compexity-based approaches to this problem. I describe very recent theoretical work of de Fontnouvelle here. de Fontnouvelle (2000) builds on Brock and Hommes (1997) to build an evolutionarily dynamic version of the 'Grossman-Stiglitz information paradox' in the context of noisy rational expectations models where agents can expend resources to obtain more accurate 'signals' of future prices, but the market 'leaks' these efforts, allowing other agents who do not pay for better signals to free ride on the efforts of those who do. This tension between information gatherers and information parasites creates a layer of dynamics across signal types (more precise signal types must pay more for their signals but earn higher gross profits from trading on those signals) that generates patterns of returns and trading volume that looks rather like real data on returns and trading volume. I.e. the EMH is accepted by simulated returns data (as it tends to be in real markets) from his model and trading volume not only fails to 'dry up' (as in Sargent's (1993) discussion of 'no-trade' theorems under rational expectations and common knowledge) but is highly persistent (as it is in the data from real markets).

Sixth, while the complexity-based approaches to 'backing off' from fully rational expectations are exciting there is a problem in taming the number of free parameters that the plethora of competing expectational schemes (also constantly evolving in the evolutionary computational approaches) brought to the analysis. As Sargent (1993) argues, economists are in the market for theories that reduce the number of free parameters not theories that increase them. Brock's paper in SFI (II) reviews recent work of Brock, Hommes, and de Fontnouvelle that adapts methods inspired by large system limits in statistical mechanics to reduce the number of parameters by recognizing that equilibrium conditions in asset markets contain terms from the heterogeneity of agents that resemble sample moments. If one takes the number of agents to infinity these 'sample moments' converge to population moments. If the 'building blocks' of the agent types come from parsimoniously parameterized 'building block' distributions then the number of free parameters is reduced from essentially the number of agents (huge) to the number of parameters in the building block distributions from which agents are built out of evolutionarily adaptive characteristics. This theory, which is still very much under construction looks rather like large economy limit theory in general equilibrium theory (cf. writers such as Aumann, Debreu, Hildenbrand, Kirman, *et al.*).

Seventh, the evolutive signal access models of de Fontnouvelle (2000) might be generalized to get 'use it and lose it' type results where better signals are purchased from 'policy insiders' for a fee, are traded on and generate superior profits as policy unfolds which impacts the potency of policy if its potency depends upon the policy makers being able to move faster than the agents or having better information than the agents. This might allow a better understanding of where policy invariance results are likely to be a problem in practice. It also may help identify channels where policy benefits are likely to be greater than its costs, especially when correction for practical problems such as political intervention and rent seeking is done.

## 13.6 Summary

This chapter has attempted to bring to the reader's attention some 'complexity-based' tools and has attempted to assess whether they might add some value to discussion of issues raised at the Siena School of Summer 1998. These issues included (i) explaining the wide differences in growth performance across regions, countries, and periods; (ii) identification of 'spillover' effects in growth contexts; (iii) and the desirability of 'backing off' from fully rational expectations modeling.

We addressed these issues by first setting out a little bit of optimal timing theory and applying it to uncovering forces that lead to faster adoption of RCR's and faster development of new markets. We then used this framework to set out five sample industrial and regulatory settings to stress how to go point-by-point through how the structure of incentives (but not all of Harberger's 1001 ways) impact optimal adoption times. Much was said about the many things that government can do to help or to hurt. After spending a fair amount of time setting out this context, we turned to a discussion of channels for potential spillovers. Then we turned to discussion of the basic econometric problem, following Manski's paper in SFI (II), of identifying spillovers in growth regressions. We suggested some ways based on Brock and Durlauf (1999) to deal with this problem. We then turned to a sketch of econometrics of hazard function estimation with spillover effects following Brock and Durlauf (1999) which is close to the spirit of the optimal timing theory developed above.

The bottom line is this. Complexity-based methods offer stimulating suggestions of potentially useful research strategies for work in the area of cycles and growth.

## References

Acar, E., Satchell, S. (eds), (1998) *Advanced Trading Rules*, Butterworth-Heineman: Oxford.

Aghion, P., Howitt, P. (1998) *Endogenous Growth Theory*, MIT Press: Cambridge, Mass.

Akdeniz, L., Dechert, W., (1997), 'Do CAPM Results Hold in a Dynamic Economy? A Numerical Analysis,' *Journal of Economic Dynamics and Control*, 21, 981–1003.

Alesina, A. (1997) 'Institutions for Fiscal Stability,' *NBER Reporter*, Winter, 1997/8.

Allen, C., Forys, E., Holling, C. (1998) 'Body Mass Patterns Predict Invasions and Extinctions in Transforming Landscapes,' Department of Zoology, University of Florida.

Altug, S., Labadie, P. (1994) *Dynamic Choice and Asset Markets*, Academic Press: New York.

Altug, S., Ashley, R., Patterson, D. (1999) 'Are Technology Shocks Nonlinear?' *Macroeconomic Dynamics* 3(4), 506–33..

Amendola, M., Gaffard, J. (1998) *Out of Equilibrium*, Clarendon Press, Oxford: Oxford.

Anderson, P., Arrow, K., Pines, D. (eds) (1988) *The Economy as an Evolving Complex System*, Addison-Wesley: Redwood City, California.

Arthur, W., Durlauf, S., Lane, D. (eds) (1997) *The Economy as an Evolving Complex System: II*, Addison-Wesley: Redwood City, California.

Baak, S. (1999) 'Tests for Bounded Rationality with a Linear Dynamic Model Distorted by Heterogeneous Expectations,' *Journal of Economic Dynamics and Control*, 23(9–10), 1517–43.

Barro, R., (1997), *Determinants of economic growth*, MIT Press: Cambridge, Massachusetts.

Barro, R., Sala-i-Martin, X. (1995) *Economic Growth*, McGraw-Hill: New York.

Baumol, W. (2001) 'Towards the microeconomics of innovation: growth engine of market economics', chapter 11 of this volume.

Black, F. (1995) *Exploring General Equilibrium*, MIT Press: Cambridge, Mass.

Böhm, B., Punzo, L. (2001) 'Productivity–investment fluctuations and structural change,' Chapter 3 of this volume.

Brock, W. (1993) 'Pathways to Randomness in the Economy: Emergent Nonlinearity and Chaos in Economics and Finance,' *Estudios Economicos*, 8, 1, 3–55. reprinted in Dechert, W. (1996).

Brock, W. (1998) 'Website: http://www.ssc.wisc.edu/ wbrock', Department of Economics, The University of Wisconsin, Madison.

Brock, W. (2000) 'Whither Nonlinear?' *Journal of Economic Dynamics and Control*, 24, 663–78.

Brock, W., Dechert, W., Scheinkman, J., LeBaron, B. (1996) 'A Test for Independence Based Upon the Correlation Dimension,' *Econometric Reviews*, 15, (3), 197–235.

Brock, W., Durlauf, S. (1995[2001]) 'Discrete Choice with Social Interactions,' NBER W.P. 5291. [*Review of Economic Studies*, forthcoming].

Brock, W., Durlauf, S. (1999) *Interactions-Based Models*, in J. Heckman and E. Leamer, (eds), *Handbook of Econometrics*, North Holland: Amsterdam.

Brock, W., Hommes, C. (1997) 'A Rational Route to Randomness,' *Econometrica*, 65, (5), 1059–95.

Brock, W., Hommes, C. (1998) 'Heterogeneous Beliefs and Routes to Chaos in a Simple Asset Pricing Model,' *Journal of Economic Dynamics and Control*, 22, 1235–74.

Brock, W., Lakonishok, J., LeBaron, B. (1992) 'Simple Technical Trading Rules and the Stochastic Properties of Stock Returns,' *Journal of Finance*, XLVII, (5), December, 1731–64.

Brock, W., Rothschild, M. (1986) 'Comparative Statics for Multidimensional Optimal Stopping Problems,' in Sonnenschein, H., (ed.), *Models of Economic Dynamics*, Springer-Verlag *Lecture Notes in Economics and Mathematical Systems*, 214, Springer-Verlag: Berlin, 124–38.

Brock, W., Rothschild, M., Stiglitz, J. (1989) 'Stochastic Capital Theory,' in Feiwel, G., (ed.) (1989), *Joan Robinson and Modern Economic Theory*, New York University Press: Washington Square, New York.

Carlson, G., Zilberman, D., Miranowski, J. (1993) *Agricultural and Environmental Resource Economics*, Oxford University Press: Oxford.

Chavas, J. (1999) 'On the Economic Rationality of Market Participants: The Case of Expectations in the US Pork Market,' *Journal of Agricultural and Resource Economics*, 24, 19–37.

Clark, W. (1985) 'Scales of Climate Impacts,' *Climate Change*, 7, 5–27.

Coleman, J. (1990) *Foundations of Social Theory*, Harvard University Press: Cambridge, Massachusetts.

Day, R. (1997) *Complex Economic Dynamics*, MIT Press: Cambridge, Mass.

Day, R., and Chen, P. (eds) (1993) *Nonlinear Dynamics and Evolutionary Economics*, Oxford University Press: Oxford.

Day, R. and Pavlov, O. V. (2001) 'Qualitative dynamics and evolution in the very long run,' Chapter 4 of this volume.

Dechert, W. (ed.) (1996) *Chaos Theory in Economics: Methods, Models and Evidence*, Edward Elgar: Cheltenham.

Dechert, W. (1998) 'Website: http://dechert.econ.uh.edu', Department of Economics, The University of Houston.

de Fontnouvelle, P. (2000) 'Informational Strategies in Financial Markets: The Implications for Volatility and Trading Volume Dynamics,' *Macroeconomic Dynamics*, 4(2).

DeLong, B. (1998) 'Website: http://econ161.berkeley.edu/', Department of Economics, The University of California, Berkeley.

Dixit, A., Pindyck, R. (1994) *Investment Under Uncertainty*, Princeton University Press: Princeton.

Dosi, G. (2000) *Innovation, Organization and Economic Dynamics*, Edward Elgar: Cheltenham.

Duffie, D. (1988) *Security Markets: Stochastic Models*, Academic Press: New York.

Durlauf, S. (1993) 'Nonergodic Economic Growth,' *Review of Economic Studies*, 60, (2), 349–66, April.

Durlauf, S., Quah, D., (1999) 'The New Empirics of Economic Growth,' *Handbook of Macroeconomics* North Holland: Amsterdam.

Evans, D., (ed.) (1983) *Breaking Up Bell*, North Holland: Amsterdam.

Evans, L., Grimes, A., Wilkinson, B., with Teece, D., (1996) 'Economic Reform in New Zealand 1984–95: The Pursuit of Efficiency,' *Journal of Economic Literature*, XXXIV, (4), December, 1856–1902.

Flaschel, P., (2001) 'Disequilbrium growth in monetary economics: basic components and the KMG working model,' Chapter 6 of this wolume.

Glaeser, E., (1997a) 'The Economics of Cities,' *NBER Reporter*, Winter 1997/8.

Glaeser, E., (1997b) 'Learning in Cities,' Harvard Insitute of Economic Research, D.P. Number 1814.

Granger, C., Terasvirta, T. (1993) *Modelling Nonlinear Economic Relationships*, Oxford University Press: Oxford.

Gunderson, L., Holling, C., Light, S. (eds) (1995) *Barriers and Bridges: To The Renewal of Ecosystems and Institutions*, Columbia University Press: New York.

Hahn, F. (2001) 'The "exogenous" in "endogenous" growth theory,' Chapter 15 of this volume.

Hahn, F., Solow, R. (1995) *A Critical Essay on Modern Macroeconomic Theory*, MIT Press: Cambridge, Mass.

Hall, R. (1996) 'Economic Fluctuations and Growth,' *NBER Reporter*, Autumn.

Hansen, L., Sargent, T. (1991) *Rational Expectations Econometrics*, Westview Press: Boulder, Colorado.

Harberger, A. (1998) 'A Vision of the Growth Process,' *American Economic Review*, 88, (1), 1–32.

Heckman, J., Singer, B. (eds) (1985) *Longitudinal Analysis of Labor Market Data*, Cambridge University Press: Cambridge, UK.

Horvath, M. (1998) 'Business Cycles and the Failure of Marginal Firms,' Department of Economics, Stanford University.

Iwai, K. (2001) 'Schumpeterian dynamics: a disequilibrium theory of long run profits,' Chapter 7 of this volume.

Jog, V., Schaller, H. (1994) 'Finance Constraints and Asset Pricing: Evidence on Mean Reversion,' *Journal of Empirical Finance*, 1, (2), January, 193–209.

Jovanonic, B., Lach, S. (1997) 'Product Innovation and the Business Cycle,' *International Economic Review*, 38, (1), February, 3–22.

Judd, K. (1998) *Numerical Methods in Economics*, MIT Press: Cambridge, Mass.

Kreps, D., Wallis, D. (eds) (1997) *Advances in Economics and Econometrics: Theory and Applications*, Cambridge University Press: Cambridge.

Krugman, P. (1996) *The Self-Organizing Economy*, Basil Blackwell: Oxford.

Kurz, M. (ed.) (1997) *Endogenous Economic Fluctuations: Studies in the Theory of Rational Beliefs*, Springer-Verlag: New York.

LeBaron, B. (1998) Website: http://www.unet.brandeis.edu/ blebaron, Department of Economics, Brandeis University.

LeBaron, B. (2000) 'Agent Based Computational Finance: Suggested Readings and Early Research,' *Journal of Economic Dynamics and Control*, 24, 679–702.

Malliaris, A. (1982) *Stochastic Methods in Economics and Finance*, North Holland: Amsterdam.

Mankiw, N., Romer, D., Weil, D. (1992) ‘A Contribution to the Empirics of Economic Growth’, *Quarterly Journal of Economics*, CVII, 407–37.

Olson, M. (1996) ‘Big Bills Left on the Sidewalk: Why Some Nations are Rich, and Others Poor,’ *Journal of Economic Perspectives*, 10, (2), Spring, 3–24.

Pesaran, M., Potter, S. (eds) (1992) *Nonlinear Dynamics and Econometrics, Journal of Applied Econometrics*, Special Issue, Volume 7, Supplement, reprinted as a book by John Wiley and Sons: New York.

Potter, S. (1995) ‘A Nonlinear Approach to US GNP,’ *Journal of Applied Econometrics*, 10, 109–25.

Putnam, R. (1993) *Making Democracy Work: Civic Traditions in Modern Italy*, Princeton University Press: Princeton.

Prescott, E. (1998) ‘Needed: A Theory of Total Factor Productivity,’ *International Economic Review*, 30, (3), August, 525–90.

Romer, P. (1998) ‘Website: http://www-leland.stanford.edu/ promer’ Graduate School of Business, Stanford University.

Rosser, J. (1991) *From Catastrophe to Chaos: A General Theory of Economic Discontinuities*, Kluwer Academic Publishers: Dordrecht, The Netherlands.

Sargent, T. (1993) *Bounded Rationality in Macroeconomics*, Oxford University Press: Oxford.

Scheinkman, J., Woodford, M. (1994) ‘Self-Organized Criticality and Economic Fluctuations,’ *American Economic Review*, 84 (May) 417–21.

Solomou, S. (1998) *Economic Cycles: Long Cycles and Business Cycles Since 1870*, Manchester University Press: Manchester.

Sullivan, R., Timmerman, A., White, H. (1999) ‘Data-Snooping, Technical Trading Rule Performance, and the Bootstrap,’ *Journal of Finance*, 54, 1647–92.

Turnovsky, S. (1995) *Methods of Macroeconomic Dynamics*, MIT Press: Cambridge, Mass.

Weidlich, W. (1991) ‘Physics and Social Science: The Approach of Synergetics,’ *Physics Reports*, 204, (1), May, 2–163.

Wheeler, C. (1998) PhD Thesis, Department of Economics, The University of Wisconsin, Madison.

# 14 Information, complexity and the MDL principle

*Jorma Rissanen*

## 14.1 Introduction

In intuitive terms, the objective of statistical modeling is to separate a given data sequence into useful learnable information and the rest, which may be viewed just as noninformative noise. The difficulty is in the formalization of the two constituents: the 'useful information' and the 'noise'. Traditionally, modeling is done by envoking a metaphysical 'true' data generating distribution, which is to be estimated from the data by minimization of an appropriate mean performance criterion, which itself is to be estimated from the data. Since the basic issue of how to formalize the useful information and the noise is not addressed such an approach cannot provide a rational explanation of why the best approximation of the 'truth' is not the most complex model fitted to the data. To avoid this disastrous conclusion one has to add an *ad hoc* term to the criterion to penalize the model complexity. But because the added term lacks any deeper meaning it does not reflect adequately the model complexity nor its effect to models' performance, and such a metaphysical assumption does not provide a sound basis for a fruitful theory of modeling.

There are alternative purely empirical approaches to model building, which are based on reuse of the available data, such as in cross-validation and bootstrapping. In these a desired data dependent performance criterion is minimized in a portion of the data and the result tested either on the rest or on a reshuffled version of the data. Although computationally demanding such approaches may work reasonably well in individual cases, but, being empirical processes, they cannot serve as a basis for a theory of modeling.

It seems to us that no solid theory of modeling can be achieved without a formal definition of 'information' in a data sequence, which can be done by a generalization of Shannon's information measure along the lines of the algorithmic theory of complexity (Hansen and Yu 1998), Kolmogorov 1965). The basic concepts have undergone a somewhat tortuous evolution, (Rissanen 1986, 1987, 1996), and in this chapter we outline some of the most recent developments (Barron *et al* 1998, Balasubramanian 1996, Grünwald 1998, Rissanen 2000a, 2000b). Briefly, the *stochastic complexity* of a data sequence $x^n = x_1, \ldots, x_n$ (or $(y^n, x^n) = (y_1, x_1), \ldots, (y_n, x_n)$

as in regression problems), relative to a class of parametric probability models $\mathcal{M} = \{P(x^n;\theta)\}$, is defined to be the shortest code length with which the data $x^n$ can be encoded, when advantage is taken of the models. Such a code length may be identified with the negative logarithm of a probability distribution $P(x^n;\mathcal{M})$, which factors as follows

$$P(x^n;\mathcal{M}) = H(x^n|\hat{\theta}(x^n))Q(\hat{\theta}(x^n)), \tag{14.1}$$

where the first factor has no useful information about the data while the second factor, describing the maximum likelihood estimate $\hat{\theta}(x^n)$, has it all. To such a factorization there often corresponds a decomposition of the data of the kind $x^n = \hat{x}^n + e^n$, where $e^n$ is the 'noise' part that cannot be compressed with the given models, and $\hat{x}^n = F(x^n;\hat{\theta}(x^n))$ is the purely predictable part. We define the code length for encoding $\hat{x}^n$ the amount of *information* in the data $x^n$ that can be retrieved with the model class $\mathcal{M}$. A similar factorization and decomposition exist also for data of type $(y^n,x^n) = (y_1,x_1),\ldots,(y_n,x_n)$, where $y^n$, given another data sequence $x^n$, is to be modeled.

We may then state that the objective in building models for data is to obtain a decomposition of the kind described, which so far has been done only for probabilistic models. Hence the goodness of any of the numerous model selection criteria proposed in the literature can be judged on how well they accomplish such a decomposition. Also, principles like Occam's razor and its variants express only the obvious: redundancy cannot add to the useful information, and since such a model cannot achieve the stochastic complexity it should be removed. The reader will recognize the factorization (14.1) as a generalization of the ordinary sufficient statistics factorization, which includes parameters, to a *universal* sufficient statistics factorization, which has no parameters. This, in turn, is an analog of Kolmogorov's sufficient statistics decomposition in the algorithmic theory of complexity (Cover and Thomas 1991) . A perfect separation of the noise from the useful information is possible only for special model classes, but asymptotically it can be done for all the usual model types. A perfect nonasymptotic separation exists in the linear quadratic regression case, (Rissanen 2000a), which we describe below.

## 14.2 Models

We begin with a brief discussion of the formal definition of a *model*. We consider sets $X$ and $Y$ and their cartesian product $X \times Y$ together with the extension $X^n \times Y^n$ to sets of strings of length $n$. Perhaps the most common type of model is defined by a function $F : X^n \to Y^n$, together with an error function $\delta(y^n,F(x^n))$, for which we take the logarithmic one $\delta(y^n,F(x^n)) = -\log f(y^n|\hat{x}^n)$, where $\hat{x}^n = F(x^n)$ and $f(y^n|\hat{x}^n)$ is a conditional density function. Many of the usual error functions, above all the quadratic one, define conditional density functions. Of particular interest to us are the parameteric models, $\mathcal{M}_\gamma = \{f(y^n|x^n;\gamma,\theta)\}$, where $\gamma$ is a structure index, such as the pair of orders $p,q$ in ARMA models, and $\theta = \theta_1,\ldots,\theta_k$ ranges over some, usually compact, subset of the $k$-dimensional euclidean space, $k$ depending on $\gamma$, such as $k = p + q$ in the ARMA models. Put $\mathcal{M} = \bigcup_\gamma \mathcal{M}_\gamma$, an example of which is the set of all ARMA models. The parameters $\theta$ often include both parameters in the function $F$ and others in the conditional probability or density function, such as the variance. Finally, an important special case is the one where the data sequence $x^n$ is absent, in which case we write $x^n$ rather than $y^n$ for the single data sequence, and $f(x^n; \gamma, \theta)$ for the models. The theory is similar for both types of data.

## 14.3 Two minmax problems

As indicated in the Introduction the key to the desired factorization and decomposition is the shortest (ideal) code length with which the data sequence $x^n$ can be encoded, when the codes are somehow designed with the models in the classes $\mathcal{M}_\gamma = \{f(x^n; \gamma, \theta)\}$ and $\mathcal{M} = \bigcup_\gamma \mathcal{M}_\gamma$, respectively. The word 'ideal' refers to the convenient habit of dropping the requirement that a real code length must be integer-valued, and hence the negative logarithm of any probability or density is regarded as an ideal code length. By a harmless abuse of notation even the word 'ideal' is often dropped. For those unfamiliar with coding we add that all these conventions are justified, because we can design codes for a large set of objects such that the real integer-length code lengths differ from the ideals by a negligible amount. And since our objects are typically the sequences $x^n$ and real-valued parameters, their sets are certainly 'large'. Whenever we need to consider code lengths for small sets of objects, such as structure indices, we use the real code lengths or their accurate estimates. To summarize, a probability distribution, a model, and a code can be identified.

With these agreements the shortest code length, relative to the class $\mathcal{M}_\gamma$, we are searching for will be of the form $-\log f(x^n; \gamma)$, which means that we are looking for a model that is universal for the class $\mathcal{M}_\gamma$ in question. The very best code length we could hope to get would, of course, be $\min_\theta \log 1/f(x^n; \theta, \gamma)$, obtainable with the ML (maximum likelihood) estimate $\hat{\theta}(x^n)$, but $f(x^n; \hat{\theta}(x^n), \gamma)$ is not a valid model, because its integral exceeds unity. This suggests the following minmax problem (Barron *et al.* 1998),

$$\min_q \max_\theta E_\theta \log \frac{f(X^n; \hat{\theta}(X^n), \gamma)}{q(X^n)}, \tag{14.2}$$

where $q(x^n)$ is any density function and the expectation is with respect to $f(x^n; \theta, \gamma)$.

It is clear that the models in the class $\mathcal{M}_\gamma$ cannot express all the statistical properties in real world data sequences $x^n$, no matter how large $n$ is. For instance, if we generate the data with a density function $g(x^n)$ which is outside the class, the data will have statistical properties different from those expressible with the models in the class. Notice, that we do not want to make the claim that the data are a sample from any distribution. Rather, we are simply using density functions to describe statistical properties in the data. This suggests that we should generalize the minmax problem (14.2) as follows

$$\min_q \max_{g \in G} E_g \log \frac{f(X^n; \hat{\theta}(X^n), \gamma)}{q(X^n)}, \tag{14.3}$$

where $G$ is a class larger than $\mathcal{M}_\gamma$. In fact, we can let $G$ consist of all distributions such that $G = \{g : E_g \log(g(X^n)/f(X^n; \hat{\theta}(X^n), \gamma) < \infty\}$. This excludes the singular distributions, which clearly do not restrict the data in any manner and hence do not specify any properties in them. Also, both the minimum and the maximum will be reached.

**Theorem 14.1** *If $\Omega$ is such that the integral*

$$C_n(\gamma) = \int_{\hat{\theta}(y^n) \in \Omega} f(y^n; \hat{\theta}(y^n), \gamma) dy^n \tag{14.4}$$

*is finite, the solution to the minmax problem (14.3) is the universal* NML *(normalized maximum likelihood) model*

$$\hat{f}(x^n; \gamma) = \frac{f(x^n; \hat{\theta}(x^n), \gamma)}{C_n(\gamma)}. \tag{14.5}$$

*Clearly,*

$$E_g \log \frac{f(X^n; \hat{\theta}(X^n), \gamma)}{\hat{f}(X^n; \gamma)} = \log C_n(\gamma) \tag{14.6}$$

*for all $g$.*

The proof is given in Rissanen 2000b.

Interestingly, for discrete data the solution $\hat{P}(x^n; \gamma)$ also solves Shtarkov's minmax problem (Shtarkov 1987),

$$\min_Q \max_{x^n} \log \frac{P(x^n; \hat{\theta}(x^n), \gamma)}{Q(x^n)} = \log C_n(\gamma). \tag{14.7}$$

Notice the important fact that the best model $\hat{f}$ involves only the models in the class $\mathcal{M}_\gamma$. Under certain conditions, satisfied for the classes of exponential distributions (Rissanen 1996),

$$\log C_n(\gamma) = \frac{k}{2} \log \frac{n}{2\pi} + \log \int_\Omega \sqrt{|I(\theta)|} \mathrm{d}\theta + o(1), \tag{14.8}$$

where one of the assumptions requires the convergence

$$-n^{-1}\{E \frac{\partial^2 \ln f(X^n|\theta)}{\partial\theta_i \partial\theta_j}\} \to I(\theta),$$

to the Fisher information matrix

$$I(\theta) = \{-E_\theta \frac{\partial^2 \log f(X|\theta)}{\partial\theta_i \partial\theta_j}\}.$$

There is another related minmax problem, originally defined for universal coding

$$\min_q \max_\theta E_\theta \log \frac{f(X^n;\theta,\gamma)}{q(X^n)}, \tag{14.9}$$

where $q(x^n)$ is any density function and the expectation is with respect to $f(x^n;\theta,\gamma)$. The minimizing model turns out to be given by a mixture

$$f_{\bar{w}}(x^n;\gamma) = \int \bar{w}(\theta) f(x^n;\theta,\gamma)\mathrm{d}\theta, \tag{14.10}$$

where $\bar{w}$ is for many model classes approximately given by Jeffreys' prior

$$\bar{w}(\theta) = \frac{\sqrt{|I(\theta)|}}{\int_\Omega \sqrt{|I(\eta)|}\mathrm{d}\eta}. \tag{14.11}$$

Moreover, the minmax value is the so-called capacity of the channel $\Theta \to X^n$, (Merhav and Feder 1995, Clarke and Barron 1990) which for iid models satisfies

$$K_n(\gamma) = C_n(\gamma) e^{-k/2+o(1)}. \tag{14.12}$$

One can show that $-\log \hat{f}(x^n;\gamma)$ and $-\log f_{\bar{w}}(x^n;\gamma)$ behave similarly for large *n*.

## 14.4 Complexity and information

If we restrict the data generating model class to the same on which the codes are designed, $G = \mathcal{M}_\gamma$, then for all codes the bound $\log C_n(\gamma)$ in (14.6) is in essence a lower bound for all models $g = f(x^n;\theta,\gamma)$ (and not only for the worst case model) except when $\theta$ ranges over a set whose volume goes to zero as *n* grows to infinity (Rissanen 1986). A generalization of this is given in Rissanen 2000b. Accordingly we are justified to define the ideal code length

$$-\log \hat{f}(x^n;\gamma) = -\log f(x^n;\hat{\theta}(x^n),\gamma) + \log C_n(\gamma) \tag{14.13}$$

to be the *stochastic complexity* of the data string $x^n$, relative to the model class $\mathcal{M}_\gamma$. We can rewrite it as

$$-\log \hat{f}(x^n;\gamma) = -\log f(x^n|\hat{\theta}(x^n),\gamma) - \log \pi(\hat{\theta}(x^n)) \tag{14.14}$$

$$f(x^n|\hat{\theta}(x^n),\gamma) = \frac{f(x^n;\hat{\theta}(x^n),\gamma)}{g(\hat{\theta}(x^n))} \tag{14.15}$$

$$\pi(\hat{\theta}(x^n)) = \frac{g(\hat{\theta}(x^n))}{C_n(\gamma)}, \tag{14.16}$$

where $g(\hat{\theta}(x^n))$ is a density function induced by $f(x^n;\hat{\theta}(x^n),\gamma)$.

For sequences $y^n$ such that $\hat{\theta}(y^n) = \hat{\theta}(x^n)$ the code lengths $-\log f(y^n;\hat{\theta}(y^n),\gamma)$, divided by *n* are virtually equal for large *n*, which means that we cannot compress the first term in (14.14) further with the given model class; it is the length of a code defining just incompressible noise having virtually no information to add to the regular learnable features that are in the optimal model $f(y^n; \hat{\theta}(x^n), \gamma)$ as defined by the second term. We call it the amount of *information* in the data sequence $x^n$ that can be extracted with the given model class. It has also been called the 'model cost'.

It was shown by Balasubramanian (1996) using quite elegant arguments of differential geometry without any appeal to the code length that one can derive a continuum of distinguishable distributions such that their equal distance lattice in the natural metric defined by the quadratic form $\theta' I(\theta)\theta$ for data sequences of length *n* has the number of elements given by $C_n(\gamma)$, (14.8). Its logarithm is called there the *geometric complexity*. Also the *NML* density function itself $\hat{f}(x^n; \gamma)$ has an interpretation in terms of differential geometry. Myung *et al.* (2000) also has illuminating examples showing that the *MDL* criterion (discussed in the next section) resulting from the minimization of the stochastic complexity (14.13) and having the important structure dependent terms gives results, especially for small amounts of data, which are superior to the common criteria, where the model complexity is penalized only through the number of parameters.

The ML estimate $\hat{\theta}(x^n)$ and the optimal model it defines have further properties to justify their fundamental importance. Let $\theta_g$ define the model that is nearest to a density function $g$ in or outside the model class $\mathcal{M}_\gamma$ in the KL distance:

$$\min_{\theta \in \Omega} E_g \log \frac{g(X^n)}{f(X^n; \theta, \gamma)} = D(g(X^n) \| f(X^n; \theta_g, \gamma)). \tag{14.17}$$

Then at least for iid models such that the parameters range over a compact set $\Omega$ the ML estimates $\hat{\theta}(x^n, \gamma)$ and $-(1/n)\log f(x^n; \hat{\theta}(x^n), \gamma)$ can be shown to converge to $\theta_g$ and $-E_g \log f(X; \theta_g, \gamma)$, respectively, with $g$-probability one (Grünwald 1998, White 1994).

## 14.5 The MDL principle

The fundamental properties of the ML-estimates and the associated measures of complexity and useful information in a data sequence are valid for each model class $\mathcal{M}_\gamma$. The model selection problem refers to finding the best structure index $\gamma$. With the same arguments as in the preceding section this amounts to finding the shortest code length for the data sequence $x^n$, relative to the model class $\mathcal{M} = \bigcup_\gamma \mathcal{M}_\gamma$, which again leads to a decomposition of the data into the useful information; i.e. the optimal model, and the incompressible noise, relative to the class $\mathcal{M}$. The criterion to find the shortest code length is the *MDL* principle, which seeks to find the index $\hat{\gamma}(x^n)$ that minimizes the stochastic complexity, or

$$\min_{\gamma} \{-\log f(x^n; \hat{\theta}(x^n), \gamma) + \log C_n(\gamma)\}. \tag{14.18}$$

The justification of this principle, as we discussed above, is the fact that it achieves the desired decomposition of the data into the information bearing part and the noise. We'll illustrate that in the concrete example of linear quadratic regression problem in the next section. Our thesis then is that the *MDL* principle and the *NML* universal model capture *all* the regular features in the data sequence $x^n$ that can be described by the model classes in question.

We conlude this section with a discussion of an inherent problem with Akaike's *AIC* criterion and how to replace it with another that has no such difficulty. In the early 1970s, Akaike (1974) set out to obtain a model selection criterion by asking for the model in a parametric class which is closest to a data generating model in KL-distance lying outside of the parametric class. This then, by less than convincing arguments aimed at estimating the distance, led into his criterion *AIC*. That there is a problem in the arguments is best illustrated by the fact that if we consider a class of nested subclasses, such as the set of polynomials of all orders $k$, where $k > m$ implies that $f(x^n; \hat{\theta}^k) \geq f(x^n; \hat{\theta}^m)$, and let the data generating model be a nonparametric one, then the nearest model is either the one with the maximum number of parameters or it does not exist at all. Also *AIC* fails to recover the data generating model even in the case where it lies within the model class.

The difficulty can be avoided if we consider the following problem

$$\min_q \max_g E_g \log \frac{\hat{f}(X^n; \hat{\gamma}(X^n))}{q(X^n)}, \tag{14.19}$$

which gives the solution

$$\hat{f}(x^n) = \frac{\hat{f}(x^n; \hat{\gamma}(x^n))}{\mathcal{C}_n},$$

where $\mathcal{C}_n$ is a normalizing constant not depending on the *MDL* estimate $\hat{\gamma}(x^n)$ of the structure parameter. Hence there is no need to evaluate the normalizing constant nor to add its code length to the *MDL* criterion (14.18).

## 14.6 Linear quadratic regression

Because of the term $o(1)$ in (14.8) we cannot calculate the normalizing constant and hence the stochastic complexity (14.13) to any desired precision for general model classes. However, the important normal family is an exception, and for the Gaussian linear regression problem the *NML* universal density function can be calculated to any precision even for small samples, and we get a complete decomposition of the data into the optimal model and an incompressible noise. This also delivers the information bearing signal sought for in the so-called denoising problem. We outline the recent findings on these developments (Rissanen 2000b) .

In the linear regression problem the data consist of pairs $(y_t, x_{1t}, x_{2t}, \ldots, x_{mt}) = (y_t, \underline{x}_t)$ for $t = 1, 2, \ldots, n$, which we also write as $(y^n, \underline{x}^n)$, and we wish to learn how the values $x_{it}, i = 1, 2, \ldots, m$, of the *regressor* variables influence the corresponding values $y_t$ of the *regression* variable. We fit a linear model of type

$$y_t = \beta' \underline{x}_t + \epsilon_t = \sum_{i \in \gamma} \beta_i x_{it} + \epsilon_t, \tag{14.20}$$

where $\gamma = \{i_1, \ldots, i_k\}$ denotes a subset of the indices of the regressor variables; the prime denotes transposition, and for the computation of the required code lengths the deviations $\epsilon_t$ are modeled as samples from an iid Gaussian process of zero mean and variance $\tau = \sigma^2$, also as a parameter. In such a model the response data $y^n = y_1, \ldots, y_n$ are also normally distributed with the density function

$$f(y^n|\underline{x}^n;\gamma,\beta,\tau) = \frac{1}{(2\pi\tau)^{n/2}} e^{-\frac{1}{2\tau}\sum_t (y_t - \beta'\underline{x}_t)^2}, \tag{14.21}$$

where the indices in $\gamma$ pick out the components of the regressor variables and define the $k \times n$ matrix $X'_\gamma = \{x_{it} : i \in \gamma\}$. Write $Z_\gamma = X'_\gamma X_\gamma = n\Sigma_\gamma$, which is taken to be positive definite. The development for a while will be for a fixed $\gamma$, and we drop the subindex $\gamma$ in the matrices above as well as in the parameters. The maximum likelihood solution of the parameters is given by

$$\hat{\beta}(y^n) = Z^{-1}X'y^n \tag{14.22}$$

$$\hat{\tau}(y^n) = \frac{1}{n}\sum_t (y_t - \underline{x}'_t\hat{\beta}'(y^n))^2. \tag{14.23}$$

We next consider the *NML* density function

$$\hat{f}(y^n|\underline{x}^n;\gamma) = \frac{f(y^n|\underline{x}^n;\gamma,\hat{\beta}(y^n),\hat{\tau}(y^n))}{\int_{Y(\tau_0,R)} f(z^n|\underline{x}^n;\gamma,\hat{\beta}(z^n),\hat{\tau}(z^n))dz^n}, \tag{14.24}$$

where $y^n$ is restricted to the set

$$Y(\tau_0,R) = \{z^n : \hat{\tau}(z^n) \geq \tau_0,\, \hat{\beta}'(z^n)\Sigma\hat{\beta}(z^n) \leq R\}, \tag{14.25}$$

which is to include $y^n$. In order to simplify the notations we suppress the super index $n$ in all the variables.

The numerator in Equation (14.24) has the simple form

$$f(y|\underline{x};\gamma,\hat{\beta}(y),\hat{\tau}(y)) = 1/(2\pi e\hat{\tau}(y))^{n/2}, \tag{14.26}$$

and the problem is to evaluate the integral in the denominator. We can do it by using the facts that $\hat{\beta}$ and $\hat{\tau}$ are sufficient statistics for the family of normal models given, and that they are independent by Fisher's lemma. With the notation $\theta = (\beta,\tau)$ rewrite $f(y|\underline{x};\gamma,\beta,\tau) = f(y|\underline{x};\gamma,\theta)$, and after some computations; for details see Dom 1996, where such calculations were done first without use of Fisher's lemma), we get

$$C(\tau_0,R) = \int_{Y(\tau_0,R)} f(y|\underline{x};\gamma,\hat{\theta}(y))dy \tag{14.27}$$

$$= A_{n,k}V_k\frac{2}{k}\left(\frac{R}{\tau_0}\right)^{k/2}, \tag{14.28}$$

where

$$V_k R^{k/2} = |\Sigma|^{-1/2} \frac{2\pi^{k/2}(nR)^{k/2}}{k\Gamma(k/2)}, \tag{14.29}$$

is the volume of the ellipsoid $B_R = \{\beta : \beta' \Sigma \beta \leq R\}$ and

$$A_{n,k} = \frac{|\Sigma|^{1/2}}{(2\pi)^{k/2}} \frac{(\frac{n-k}{2e})^{\frac{n-k}{2}}}{\Gamma(\frac{n-k}{2})}. \tag{14.30}$$

We then have the *NML* density function itself for $0 < k < m$

$$-\log \hat{f}(y|\underline{x};\gamma,\tau_0,R) = \frac{n}{2}\ln\hat{\tau}(y) + \frac{k}{2}\ln\frac{R}{\tau_0} - \ln\Gamma(\frac{n-k}{2}) - \ln\Gamma(\frac{k}{2}) + \ln\frac{4}{k^2} + \frac{n}{2}\ln(n\pi). \tag{14.31}$$

In order to get rid of the two parameters $R$ and $\tau_0$, which clearly affect the criterion in an essential manner, set them to the values that minimize (14.31): $R = \hat{R}(y) = n^{-1}\hat{\beta}'(y)\Sigma\hat{\beta}(y)$ and $\tau_0 = \hat{\tau}(y)$. However, the resulting $\hat{f}(y|\underline{x};\gamma,\hat{\tau}(y),\hat{R}(y))$ is not a density function. We rectify this by the same normalization process as above:

$$\hat{f}(y|\underline{x};\gamma) = \frac{\hat{f}(y|\underline{x};\gamma,\hat{\tau}(y),\hat{R}(y))}{\int_Y \hat{f}(z|\underline{x};\gamma,\hat{\tau}(z),\hat{R}(z))dz}, \tag{14.32}$$

where the range

$$Y = \{z : \tau_1 \leq \hat{\tau}(z) \leq \tau_2, R_1 \leq \hat{R}(z) \leq R_2\}$$

will be defined by four new parameters. Again the integration can be performed, (Rissanen 2000a), and the negative logarithm of $\hat{f}(y|\underline{x};\gamma)$ is given by

$$-\ln\hat{f}(y|\underline{x};\gamma) = \frac{n-k}{2}\ln\hat{\tau}(y) + \frac{k}{2}\ln\hat{R}(y) - \ln\Gamma(\frac{n-k}{2}) - \ln\Gamma(\frac{k}{2}) \tag{14.33}$$

$$+ \ln\frac{2}{k} + \frac{n}{2}\ln(n\pi) + \ln\ln\frac{\tau_2 R_2}{\tau_1 R_1}. \tag{14.34}$$

This time the last term involving the new parameters does not depend on $\gamma$ nor $k$, and we do not indicate the dependence of $\hat{f}(y|\underline{x};\gamma)$ on them. Almost the same criterion was obtained in Hanson and Yu 1998 by evaluation of a mixture density for Zelner's prior.

As a final step we wish to extend the density function $\hat{f}(y|\underline{x};\gamma)$ to the larger class of models, defined as the union over all the index sets $\gamma$, and to obtain a criterion for finding the optimal index set and the associated optimal model. We begin with the MDL estimator $\hat{\gamma}(\cdot)$, obtained by minimization of the ideal code length for the data $-\ln\hat{f}(y|\underline{x};\gamma)$ with respect to $\gamma$. Although the result $\hat{f}(y|\underline{x};\hat{\gamma}(y))$ is not a density function we get one by the normalization process

$$\hat{f}(y|\underline{x};\Omega) = \frac{\hat{f}(y|\underline{x};\hat{\gamma}(y))}{\int_{\hat{\gamma}(z)\in\Omega}\hat{f}(z|\underline{x};\hat{\gamma}(z))dz}, \tag{14.35}$$

where $\Omega$ is a set of indices such that it includes $\hat{\gamma}(y)$. The denominator, call it $C$, is given by

$$C = \sum_{\gamma \in \Omega} \hat{P}_n(\gamma), \tag{14.36}$$

where

$$\hat{P}_n(\gamma) = \int_{\{z:\hat{\gamma}(z)=\gamma\}} \hat{f}(z|\underline{x};\hat{\gamma}(z))dz. \tag{14.37}$$

In analogy with $\hat{f}(y|\underline{x};\gamma)$ we call $\hat{f}(y|\underline{x};\Omega)$ the *NML* density function for the model class with the index sets in $\Omega$, and we get the final decomposition

$$\begin{aligned} -\ln\hat{f}(y^n|\underline{x}^n;\Omega) &= \tfrac{n-\hat{k}}{2}\ln\hat{\tau}(y^n) + \tfrac{\hat{k}}{2}\ln\hat{R}(y^n) - \ln\Gamma(\tfrac{n-\hat{k}}{2}) \\ &-\ln\Gamma(\tfrac{\hat{k}}{2}) + \ln\tfrac{1}{\hat{k}} + \text{ Const }, \end{aligned} \tag{14.38}$$

where we include in Const all the terms that do not depend on the optimal index set $\hat{\gamma}$ of size $\hat{k}$. The terms other than the first define the length of a code from which the optimal normal model, defined by the ML parameters, can be decoded, while the first term represents the code length of the part of the data that adds no further information about the optimal model. It may be viewed as noise. Hence this decomposition is similar to Kolmogorov's sufficient statistics decomposition in the algorithmic theory of information, and it is also seen to extend the ordinary sufficient statistics, as defined for certain parametric families, to parameter free *universal* sufficient statistics.

By applying Stirling's approximation to the $\Gamma$-functions we get the *NML* criterion for $0 < k \leq m$

$$\min_{\gamma \in \Omega}\{(n-k)\ln\hat{\tau}(y^n) + k\ln(n\hat{R}(y^n)) + (n-k-1)\ln\frac{n}{n-k} - (k+1)\ln k\}, \tag{14.39}$$

where $k$ denotes the number of elements in $\gamma$.

It seems that in order to find the minimizing index set $\gamma$ we must search through all the subsets of the rows of the $m \times n$ regressor matrix $X'$. However, this can be avoided if we make a linear transformation $AX'$ of the regressor matrix such that $AX'XA'/n = I$. Let the new ML parameters for the maximal number of rows $k = m$ be $\hat{\alpha} = \hat{\alpha}_1, \hat{\alpha}_2, \ldots, \hat{\alpha}_m$ and $\hat{\sigma}^2$. Further, let $\hat{\alpha}^2_{(1)} \geq \hat{\alpha}^2_{(2)} \geq \ldots \geq \hat{\alpha}^2_{(m)}$ so that $(i)$ is the index of the $i$'th largest parameter in absolute value. Then the criterion (14.39) is equivalent with

$$\min_{k}\{(n-k)\ln(\hat{\alpha}'\hat{\alpha} - \hat{R}) + k\ln(n\hat{R}) + (n-k-1)\ln\frac{n}{n-k} - (k+1)\ln k\}, \tag{14.40}$$

where $\hat{R}$ is the sum of either the $k$ largest or the $k$ smallest squares $\hat{\alpha}_i^2$ (Rissanen 2000a). Hence, the optimum $\gamma$ can be found with no more than $m$ evaluations of the criterion.

For the so-called denoising problem an orthonormal regressor matrix is easily obtained with wavelets, and the criterion (14.40) provides a natural separation of noise from the data as its incompressible part, which appears to be superior to any *ad hoc* criterion; for numerical examples see Rissanen 2000a.

## 14.7 Conclusions

We have described a formal definition of *complexity* and (useful) *information* in a data sequence, as a foundation for a theory of model building. These notions can be obtained to a good approximation from a universal *NML* (Normalized Maximum Likelihood) model for parametric model classes as a solution to a minmax problem. Theorems exist which demonstrate that such a universal model provides an extension of the usual sufficient statistics decomposition to a parameter free *universal* sufficient statistics decomposition, and accomplish the desired decomposition of the data. For a collection of model classes the best can be found with the *MDL* principle. We illustrate such a decomposition and the resulting criterion for the Gaussian family in the basic linear regression problem, for which they can be computed exactly even for small amounts of data.

## References

Akaike, H. (1974), ‘Information Theory and an Extension of the Maximum Likelihood Principle,’ *Second international symposium on information theory*, ed. B.N. Petrov and F. Csaki, Akademia Kiedo, Budapest, 267–81

Balasubramanian, V. (1996)‘Statistical Inference, Occam’s Razor and Statistical Mechanics on the Space of Probability Distributions’, *Neural Computation*, 9, (2), 349–268, 1997 http://arxiv.org/list/nlin/9601

Barron, A.R., Rissanen, J. and Yu, B. (1998) ‘The MDL Principle in Modeling and Coding’, special issue of *IEEE Trans. Information Theory* to commemorate 50 years of information theory, IT-44, (6), October 1998, pp 2743–60

Clarke, B. S. and Barron, A. R. (1990) ‘Information-Theoretic Asymptotics of Bayes Methods’, *IEEE Trans. Information Theory*, IT-36, (3), 453–471, May 1990.

Cover, T. M. and Thomas, J. A. (1991) *Elements of Information Theory*, Wiley, New York,.

Dom, B. (1996) ‘MDL Estimation for Small Sample Sizes and Its Application to Linear Regression’, *IBM Research Report* RJ 10030, June 13, 1996.

Grünwald, P.D. (1998) *The Minimum Description Length Principle and reasoning under Uncertainty*, PhD thesis, Institute for Logic, Language and Computation, Universiteit van Amsterdam.

Hanson, M. and Yu, B. (1998) ‘Model Selection and the Minimum Description Length Principle’, to appear in Journal of American Statistical Association, June, 2001; see also http://cm.bell-labs.com/stat/doc/mdl.pdf

Kolmogorov, A.N. (1965) ‘Three Approaches to the Quantitative Definition of Information,’ *Problems of Information Transmission* 1, 4–7.

Li, M. and Vitanyi, P.M.P. (1997) *An Introduction to Kolmogorov Complexity and Its Applications,* second edition, Springer-Verlag, New York.

Merhav, N. and Feder, M. (1995) ‘A Strong Version of the Redundancy-Capacity Theorem of Universal Coding’, *IEEE Trans. Information Theory*, IT-41, (3), 714–22.

Myung, I., Balasubramanian, V. and Pitt, M. (2000) ‘Counting Probability Distributions: Differential Geometry and Model Selection’ *Proceedings National Academy of Science*, 97, 21 (October 2000), 11170–5.

Rissanen, J. (1986) ‘Stochastic Complexity and Modeling’, *Annals of Statistics*, 14, 1080–100

Rissanen, J. (1987) ‘Stochastic Complexity’, *J. Royal Statistical Society, Series B*, 49, (3) (with discussions) 223–65

Rissanen, J. (1996) ‘Fisher Information and Stochastic Complexity’, *IEEE Trans. Information Theory*, IT-42, (1), pp 40–7

Rissanen, J. (2000a) ‘MDL Denoising’, *IEEE Trans. on Information Theory*, IT-46, (7), November 2000. Also http://www.cs.tut.fi/ rissanen/.

Rissanen, J. (2000b) ‘Strong Optimality of Normalized ML models as Universal Codes and Information in Data’, http://www.cs.tut.fi/ rissanen/ (to appear in *IEEE Trans. Information Theory*, 2001

Shtarkov, Yu. M. (1987) ‘Universal Sequential Coding of Single Messages’, Translated from Problems of Information Transmission, 23, (3), 3–17.

White, H. (1994) *Estimation, inference, and specification analysis*, Cambridge University Press, Cambridge, UK.

# 15 The ‘exogenous’ in ‘endogenous’ growth theory

*Frank Hahn*

## 15.1 Introduction

In this chapter I examine the claim of the new growth theories that they provide an ‘endogenous account’ of economic growth. The first question is: what would count as an endogenous account? Or perhaps better: what would not count? For it is easier to answer this.

The Solow model [1956] is perhaps the paradigm of an exogenous growth model. In that model the rate of growth of Harrod-neutral technical progress is taken as given. Of course the type of technical improvement is also fixed. On the other hand economic theory is used to show that every equilibrium path of the economy seeks the steady state. Many if not most of current theories do not even ask or answer the question of convergence. Yet again Solow does not explain why we need only consider equilibrium paths (Harrod had distinguished between the ‘warranted’ and the ‘ actual’ growth rate). But the same is true of current ‘endogenous’ theories. One can continue . . . for instance the saving rate is given in Solow while it is deduced from the maximisation of a Ramsey integral of a given intertemporal utility function in many ‘endogenous’ theories.

So it seems clear that while it is easy to distinguish between endogenous and exogenous variables it is uncertain how to distinguish theories by the same criteria. On reflection one reaches the conclusion that all theories in all subjects must base themselves on some exogenously given elements. But in that case what is it about recent growth theories that entitles them to be called ‘endogenous’? It seems that the answer is simply that the exogenously given rate of technical progress has been abandoned in favour of a stochastic profit maximising use of resources to search for technical improvements. This can be combined with an intertemporal utility maximising choice of training. Since both can involve increasing returns and externalities some fairly drastic changes in received theory may result: e.g. multiple steady states and non-perfect competition. But all arguments are derived from equilibrium conditions and functional forms are chosen to ensure that some steady state equilibrium exists.

Clearly the scope of the theory has been enlarged. Has this been sufficient to earn the character of an ‘endogenous’ theory? It is time to ask what we mean by the question. In an endogenous theory all the elements which explain and determine growth are explained as an outcome of the rational decisions of agents. So one may be willing to accept rationality of agents as exogenously given. Or more ambitiously it may be deduced from less immediate postulates, e.g. a Darwinian struggle for survival. How the theory proceeds on this has important consequences for its scope; to how many societies can it be applied?

Let us stick to a less ambitious interpretation to see what is needed. Rationality of decisions (or of beliefs) by itself will not get us very far. We shall surely need to be given the preferences of agents, their distribution, their endowments and their distribution. Of course we shall need to know what information is available. In preferences one includes attitudes to uncertainty and beliefs. One also must find a way to describe the educational endowments and the production process of education. The structure of production i.e. the mix of industry and agriculture as well as the mix of competitive and imperfectly competitive activities all need to be specified or explained. So will the endowment of public goods.

The list is large but I have included in it only what seems most relevant for growth and have as yet not even mentioned an account of the manner of technical progress.[1] This will be discussed below. At this stage I hope that it will be clear why I regard the somewhat hopeless task of a general definition of ‘endogenous’ to be unnecessary. It is clear to the naked eye that functions describing the outcome of the choice of education, or the acquisition of new technological knowledge by a choice of R and D will not suffice for a model of growth in which the most important elements are themselves explained, by the model. That of course does not mean that these partially endogenous models are not valuable. But it does suggest that it would be useful for consumers of these theories to be warned that they are not being offered a theory of economic history. This chapter constitutes such a warning.

## 15.2 Schumpeterian models

In my view the clearest exposition of a possible modern Schumpeterian approach is that of Aghion and Howitt (1998). It also allows me to make some of the above points more precisely. I shall only discuss some of their simpler models but I will note the results of variation.

The idea is that output $y$ is produced according to a monotone concave production function with intermediate goods as inputs:

$$y = AF(x)$$

Innovations replace some intermediate goods by others and raise $A$. They assume that $A$ is raised by a constant factor $\gamma > 1$. Labour in turn can be allocated to research or to the production of intermediate goods. Innovations arrive at a Poisson rate $\lambda.\phi(n)$ where $\lambda$ indicates the productivity of the research technology and $n$ workers are allocated to research. This amount $(n)$ is given by the equilibrium condition

$$w_t = \lambda\phi'(n_t)V_{t+1}$$

where $t$ is the number of the innovations which have taken place and $V_{t+1}$ is the expected pay-off to the $(t+1)^{\mathrm{th}}$ innovator:

$$V_{t+1} = \pi_{t+1}/(r + \lambda\phi(n_{t+1}))$$

where $n_{t+1}$ is the amount of labour in R and D after the $(t+1)^{th}$ innovation and $\pi_{t+1}$ is the flow profit of the $(t+1)^{th}$ intermediate good monopolist. They interpret $(r + \lambda\phi(n_{t+1}))$ as the 'obsolescence adjusted interest rate' and define

$$\pi_{t+1} = A_{t+1}\hat{\pi}(w_{t+1}/A_{t+1}) \quad \text{with } \pi' < 0.$$

There must be no profit that can be made (in equilibrium) by reallocating labour between the production of intermediate goods and research. Hence

$$w_t = \lambda\phi'(n_t)\gamma\pi(w_{t+1})/\{r + \lambda\phi(n_{t+1})\}$$

Since labour can be indifferently used in research or production we obtain a second equilibrium equation for labour market clearing ($\ell^0$ = *labour* allocated to the production of $x$)

$$L = n_t + \ell^0(w_t)$$

In steady state one drops the '$t$' and has

$$w = \lambda\phi'(n)\phi\pi(w)/\{r + \lambda\phi(n)\}$$

$$n + \ell^0(w) = L$$

This can easily be shown to yield a unique steady state growth rate which will be an increasing function of $\lambda$ and $\gamma$ and $L$ and a declining one in $r$.

Schumpeter believed that competition increased the rate of innovations. If we write $F(x) = x^{\alpha}(0)$ the parameter $\alpha$ is a measure of the degree of competition and the derived demand curve faced by the monopolist has an elasticity $1/(1-\alpha)$. So now

$$\pi = wx(1-\alpha)/\alpha \equiv w(L-n)(1-\alpha)/\alpha$$

So $A$ becomes

$$1 = \lambda\phi'(h^*)\gamma(L-n^*)(1-\alpha)/\alpha//\{r + \lambda\phi(n^*)\}$$

So $n^*$ is declining in $\alpha$ and competition is bad for growth. (It is not clear that this is a faithful interpretation of Schumpeter.)

The above are the bare bones of $A.H$ and the simplest. By introducing many sectors, capital into R and D and a sequential adoption rate they can attain answers which make competition good or bad for growth depending on parameter values. They are to be praised for isolating some of the key issues which include obsolescence and the foresight of it.

It is not the lack of realism of the construction which is of interest here, but rather the question: is it an endogenous model of growth. It is clear that some of the more important parts are exogenous. Not only $\lambda$, $\gamma$ and $\alpha$ which play a considerable part but the functional forms, the postulate of universal risk neutrality, the homogeneity of the ability composition of labour, the perfect information of agents and the postulate that there is either a steady state or that markets clear at all dates. Of course even pretty closed theories depend on parameter values: for instance in particle physics, the spin of the electron. But physicists can measure these parameters, to put it no more strongly, better than we can the ones relevant to us. It is precisely our relative inability to measure that leads to exogenous variables. That and the immense complexity of an interdependent group of variables.

I hope that it will not be regarded as presumptuous if I sketch a model I published in 1990 (for a 1988 conference) to illustrate the latitude plausibility gives us in functional specification of R and D. Nowadays this approach is rather commonplace.

Unlike AH I took the production function of firm as Cobb–Douglas:

$$y_t = a_t l_t k_t$$

where $k_t$ is the capital to efficiency units of labour. I took a deterministic formulation because I did not know how to deal with risk aversion of firms.[2] Also I worked in continuous time, so that if $r_t$ is the amount of research per efficiency unit one obtained

$$\dot{a} = h(r_t)a_t$$

where $h(r) = \alpha + \lambda(r)$ and $\lambda(r)$ is logistic. I then postulated that the difficulties of innovation first decline and then increase with the amount of research already done. (Compare this with Romer [1986].)

To capture externalities I let the cost of research per efficiency unit $c(\cdot)$ be $c(r,R)$ with $R =$ *average* amount of research in the economy,

$$c_r(r,R) > 0 \text{ all } (r,R),\ c_{rr}(r,R) > 0 \text{ when } \ r \neq R,\ c_{rr} + c_{rR} = 0 \text{ if } \ r = R.$$

One can find many plausible arguments for those kinds of externalities. These include those given by AH but also the likelihood that a higher $R$ will lead to a larger community of research workers and a greater familiarity with research methods. For the rest I do not at all improve on A H. R and D is carried out inside the producing firm and there are no specialised research firms. I then use the calculus of variations to find the firm's present value profit maximising plans. To this I add the equilibrium condition of self-finance which is clearly implausible. Moreover I take an exogenous Solowian saving ratio. That too is limiting, but the model can be rectified in the Chicago direction rather easily but at the Chicago cost that savers have perfect foresight. (Actually in a deterministic setting such as here this would not be as silly as usual.) There is only one good produced.

It turns out that there are three steady states: one with no research at all so that $\alpha$ is the Solowian rate of technical progress. In all steady states $r = R$ but there is more research in one than the other. Of course in steady state all firms take $R$ as given when they decide on $r$, (and only two can be stable).

I think that I prefer the AH model with its spirit of struggle for existence and its explicit attention to obsolescence.[3] On the other hand in my version the role of *R* is also plausible. It indicates whether one lives in a world where research is a habit or not, and my formulation suggests that it may or may not be the case. This gives just a little greater flavour of endogeneity. On the other hand again I, no more than others, distinguish labour by its ability to research and produce. But my point is not realism. It was to underline how much that is taken as endogenous here really is not. This brings me to my next point.

## 15.3 Equilibrium

When we say that a variable $z$ is endogenous to a model we mostly mean that it is determinable by equilibrium conditions. In growth theory this involves expectations in an essential way. In my view they also involve risk attitudes in an essential way. For instance it seems that in nineteenth century England risk aversion was less than in Germany. One also needs to make postulates concerning the information flows and the ability to use them. So when one writes down equilibrium equations one is bringing into the story many elements not explained by any theory.

If we pay explicit attention to these then we may find our way to a theory which can indeed explain our observation – say economy A grows slowly because it has many risk-averse producers – say farmers. But it is not clear whether economists can say why this should be so.

My point as usual here is not the usual grouse concerning lack of realism. It is that when we use the canonical paradigm of economic theory we are rarely in a position to attain purely endogenous results. Or rather what seems like those are not genuinely so.

As a good example take the influential Chicago method. It needs to postulate that expectations are rational and look at rational expectations equilibria. The time horizon for these expectations is infinity. There are possible learning theories – which lead to rational expectations over finite time but, in the nature of the case, such expectations over the infinite future could not be proved. But equally seriously, almost all of the authors make no allowance for risk attitudes or, rather, postulate risk-neutrality. In our theories of course this means that utility functions are taken to be exogenous. This in my view is the right attitude to take by economists. But there is almost surely a distribution of such functions so that the more subtle hypothesis that tastes etc. are stable through infinite time needs to be invoked. This is certainly an exogenous piece of theory. For long run growth theories it does not seem persuasive. In addition Chicago always postulates perfect competition and production possibilities which allow this. All improvements in technique are either in intermediate goods or due to improved education but, as in Marshall, external to the competitive firm. This feature is not explained by the theory and hence an exogenous element. I do not know whether it is claimed that the theory fits the facts (so far), but I doubt it, since it is too aggregated to be confronted with some of the most important episodes in growth, e.g. the introduction of the steam engine, the internal combustion engine, electricity, computers, etc.

This discussion mirrors one in evolution. Gould maintains if exogenously given stochastic geological events had been different than they have been, present day creatures would not only be different but intelligence might never have arisen. The Cambridge professor of evolution believes that whatever the external geological etc. events intelligence would have evolved – I assume because of the advantage it confers. In our context this is something of a pro-Chicago argument, since it argues that competition and intelligence are almost bound to lead to growth whatever particular form it takes. But that is a good deal more modest claim. In particular it could be made without the full equilibrium paraphernalia since in this context it is bound to be ill-defined without imports of exogenous factors.

As a matter of fact there are theories which give a fully endogenous account of why a (Nash) equilibrium should occur. *But* they rarely include technical progress and, what is worse, the equilibrium is not unique. This then requires initial conditions and a more or less exogenously specified process of the evolution of strategies to get us where we want to be. A variety of these have been proposed but the outcome is more in the spirit of Gould since some processes converge to the risk-dominant rather than the Nash equilibrium.

## 15.4 Expectations again

Chicago has rational price expectations which seems somewhat odd in a world in which future technology has to be anticipated albeit stochastically. Suppose for instance that a high probability is given to a future viable electric car. The first question then is – say for oil companies – its probable date. The second is how would it affect all the prices relevant to an agent’s decision. This is an example which is not in itself hostile to endogeneity – for instance the pollution etc. of present cars is what sets off the search for electric ones. But it is, I believe, hostile to endogeneity as perpetual equilibrium.

AH – good Schumpeterians – pay special attention to anticipated obsolescence and to Schumpeterian temporary monopoly for the innovator. But I think that there is at any time a rather large array of vintages so that at best one could calculate the expected average length of anyone of them. (Like everything else, this is discussed by AH). I am sure that one can construct some kind of theoretical picture of this – indeed AH do so. But this depends on information available, (or obtainable), and on good telescopic faculties which we recall was one of the characteristics of the Schumpeterian entrepreneur. Relatively small risk-aversion was another. The proportion of these in a population at present is exogenous. (See below).

Moreover as a matter of plain fact the economies we know have not been and are not perfectly competitive. Firms then must forecast expected demand functions – their own prices they will set in the light of these forecasts. Except for the simplest cases neither I nor, I believe, any one else knows how to construct an equilibrium model consisting of these kind of firms. But even if we did it is obvious that there would be a very large exogenous – unexplained – component. In most endogenous theories new products are continually appearing and it is not clear how past experience leads one to rationally expected demand curves. One could in the usual way think of agents engaged in a game. But the game will be changing as strategy sets and payoffs are changed by ‘progress’. Again it will be possible to cobble together some hypotheses but, on the whole, it seems to me that this will replace what might have been taken as a simple exogenous process by a more fancy one which must cast doubt on, for instance, agents at all time being in a Nash equilibrium.

Returning to some remarks of AH it must be doubtful that the elasticity of demand should be treated as exogenous in an endogenous theory or that it is a good measure of monopoly. For one thing most firms produce multiple products, for another one would have thought the position of the demand curve – share of market – was also a monopoly indicator. Of course there is something to the Kalecki measure but not enough, it seems to me, to measure the extent of competition. For instance legal restrictions on market share or on getting together on pricing also matter. There probably exist, or could exist, 'endogenous' theories of how such laws arise, but we must take this as exogenous. (Recall that not so long ago kings licensed monopolies.)

## 15.5 Information

Risk-attitude and competence characterised the Schumpeterian entrepreneur. Their number Schumpeter believed to depend on the culture including the religion of the economy. Interestingly enough he thought that the 'routinisation' of R and D would not only lead to the entrepreneur's obsolescence but to a slow down in technical progress. But in any case R and D does depend essentially on people capable of generating new knowledge not only inside the R and D outfit but outside it.

The DNA revolution has transformed R and D in pharmaceutical industries and agro-businesses. But this revolution in knowledge was brought about by exceptional people the fraction of which in the population is surely at present exogenously given. It also, of course, depended on these people having the opportunity to do their revolutionising, on the transmission of this knowledge and on sufficiently educated people to recognise the importance of what has been accomplished, (see Arrow [1974]). Clearly this brings education into the picture (see below) and the whole is permeated by externalities.

There are many examples of revolutionary inventions which for a long time were ignored by those who could have profited from their use. The jet engine is one, as is the fluorescent bulb. I do not know the facts sufficiently well to make the point completely convincing: to recognise the benefit of an advance is almost as important as the advance itself. For instance it is reported (Nelson and Wright [1992] in their excellent survey) that when electronic computers first emerged to satisfy the US military, the general view, including at IBM, was that it would have few civilian uses and the invention languished for quite some time. This seems to me more than a case of mistaken expectations but rather a lack of imagination. People with that kind of imagination are rare and certainly their representation in the economy is a matter of culture and largely exogenous for the economist. It is not clear whether one could regard it as an output of education. (See below.)

But the basic point is (i) finding new information, (ii) whether and how it is transmitted and (iii) how widely it can be used. I doubt that Crick and Watson needed the incentive of monetary rewards to generate DNA information. However, there were other rather obvious incentives (which I have not seen discussed in the literature). Certainly adoption by commercial firms or even feeding new information into R and D depends on their calculation of probable rewards. These in turn, as Schumpeter argued, depend on the estimated length of money rents (and on patent law). Clearly that is part of an 'endogenous' argument, but only a part. Many of the relevant elements seem clearly exogenous.

Of course much information is available by publication and word of mouth. But because information is a public good its transmission, if it is valuable, will be imperfect. Some information when it is directly or indirectly revealed will be a surprise to agents who cannot be assumed to have already adapted to it in their plans. Except, that is, for one thing: the positive probability given to the appearance of new innovations. This aspect, 'creative destruction' is well treated by AH and of course enters growth as a negative element. But the probability of such obsolescence will vary from case to case, e.g. the biro does not, and computers do, seem to run this risk with low and high probability. Once again exogenously given risk attitudes play as large a part as probabilities. These in turn depend on the time horizon (effective discount rate) of the agents concerned.

Nelson and Wright [1992] report the predominant view of historians that the early US productivity superiority (from about 1840) to 1960 had a great deal to do with superior organisation, abundant raw material and a large domestic market. (The US for long was a high tariff economy.) None of this seems to be captured by existing models of endogenous growth.[4] Organising and managing etc. is not a matter of R and D but rather the fruit of a culture which led to the emergence of a 'correct' way to organise and manage. They emphasise that the education system did not produce many outstanding scientists who had to go to Europe to study. It is true that high real wages made general education advantageous and that a generally educated workforce helped to sustain the US system of production. But later, immigration of ill-educated workers does not seem to have led to a slow down. Certainly information was of almost a collective sort – one knew how others organised and so the usual caveats do not apply.

However after the second world war and the rise of 'science based' industries, this picture changes. In particular there was a great growth in resources devoted to R and D in the US and secrecy of certain aspects of a firm's activities became important. But it was never watertight. The rise of science based industries itself depended on the progress of science which, in turn, depended on the evaluation of the benefits it would confer. It is interesting to note that the government played a significant role in this (partly through universities, partly through subsidies).

As a description these (and other aspects in particular education) matters are very clear. How to convert this into an endogenous theory of growth is not. The very fast growth in scientific knowledge seems like an exogenous cumulative process. Once the genetic code was discovered for instance, many researchers were attracted to this field and new knowledge rapidly accumulated. This sort of thing is hard to embody persuasively in a functional form and even if it could be, it would be quite unclear whether it also embodies other scientific advances, e.g.

in superconductivity. Least clear is whether these various theoretical descriptions can be made to yield a steady state with exponential growth.

But of course post-war scientific advances, for a time, led to greater encouragement of fundamental science and the hiring of scientists by firms. So of course there is something 'endogenous' to economic models of growth. My contention is that it is only 'something'. For instance, many fundamental sciences by now receive greatly reduced subsidies from government and firms often use scientists as managers. There are still considerable risks in new science based economic ventures, for instance in bio-technology. Everyone sees ever more scientific knowledge emerging, but there is not only uncertainty as to timing but as to kind. This brings me back to risk attitudes.

## 15.6 Risk attitudes

I have proceeded on the hypothesis that these are exogenously given. This is only correct in a somewhat subtle way.

Prevailing risk attitudes of others have an effect on my own when I manage a firm. When I see others taking considerable risks with some success I shall be more inclined to do so also. Moreover the high risks undertaken by others may spell danger for me (obsolescence, etc.). This is only partly an effect on risk attitudes but on the risks I run. Here is what AH mean by the effects of competition or a higher arrival rate of new knowledge. But attitudes to risk are also likely to be affected. If the prevailing social norm is not to flinch from risk, my attitude will be different from what it would be for a more conservative social norm. If one thinks of these matters game-theoretically, there may be many outcomes (e.g. Nash equilibria) and it will be a path-dependent and so partly an exogenously given one.

I believe that it would not be too hard to put all of this formally. But I am not attempting a new theory here. My objection to most endogenous growth models is that they ignore risk attitudes. It seems to the highest degree unlikely that managers do not consider the variance (possibly also higher moments) of their expectations of future new knowledge and of their chances of producing it, say by R and D.

In production, risk attitudes of firms is related to the way they are governed. A single powerful manager is not all that common and much is decided by committees, etc., and I do not know of a good theory of the risk attitudes of such bodies. But, as I say, there are cases where the attitude of a dominant manager is determining. In the UK Lord Weinstock at GEC was such a person and he was extremely risk averse. At some time the company had two billion pounds of liquid assets and Britain never developed its own viable electronics industry. An economy where powerful magnates are the norm is not one in which attitudes to risk can be ignored and they, I believe, must to some extent at least be taken as exogenous.

But there is hope for serious endogenous theorising in the global economy. The reasons are clear – peculiarities, whether national or personal have a harder time of surviving or being important. Although there seems to be a furious debate on 'convergence', this I think cannot be denied and differences will turn more on exogenous physical peculiarities such as infra-structure etc. That is, ultimate differences will depend on these. On the way there are inherited education systems etc., all of which are exogenous.

Lucas who believes that a good way of modelling growth is by a Ramsey model and who is willing to invoke externalities (here in education), simply appeals to different steady state equilibria. But I find the analysis very unconvincing, if for no other reason than that one needs to invoke all sorts of institutional elements to argue that, say, Mozambique is on an optimum growth path. Nonetheless, as usual, I admit that the analysis evidently reveals some real features – the importance and externalities of education.

Now under the present theory of uncertainty all this may sound as if I am objecting to a particular parametisation of utility functions which we all take as exogenous. But a social historian would not do so. Exogenously given utility functions are acceptable when we are studying situations over a stretch of time which is relatively brief. But as $t$ goes to infinity that seems far-fetched. At a guess a medieval farmer was much more risk averse than a modern one. If nothing else we are looking at the curvature of $U$ at higher income levels, and it simply is implausible that it is linear over the whole set of possible outcomes. For an endogenous theory of growth we would like an endogenous theory of the evolution of risk attitudes. As far as I know we do not have one.

## 15.7 Labour and education

There have been many attempts at endogenous population theories. They are complex and contentious and most endogenous growth models (but see Jorgenson [1961]) take them as exogenous. Notoriously, of course, not Malthus and other classical economists.

Even so, if labour is measured in efficiency units some possibilities of endogeneity arise even for constant population. Indeed since economists universally regard labour as a source of disutility, the vast increase in labour productivity which has taken place must lead us to suppose that the supply of raw labour has diminished over the century – the amount of time taken in leisure has increased. Curiously enough this rather obvious manifestation of endogeneity in growth theory is rarely analysed. Here too there will be important externalities as to what counts as ‘long hours’ or even ‘hard work’ changes. In a study of exponential steady state growth this matter deserves attention since the limit of the growth of efficiency and leisure choices need to be given. (See Hahn [1990]).

But I do not propose to delve deeply into this. What has become important in endogenous theorising (ever since Lucas’s [1988] paper) has been the role of education. The choice of education is discussed in the usual marginal expected benefits and costs terms. Education itself adds to the number of efficiency units of labour directly, quite apart from its contribution to the process of innovation. But the ‘education industry’ if one may so call it is left largely unmodelled. There have been large innovations in that ‘industry’ from a concentration on the classics and theology, to mathematics and science. No doubt these changes can also be partially accounted for by the greater use business could make of numerate workers. But a good deal of these changes were due to governments and one needs a theoretical account of how these respond to the needs of business. There is also an egg and chicken problem. Until there are sufficiently trained numerate people, others cannot choose to be trained in these skills. The Lucas procedure of letting the education of labour be entirely determined by the trade-off between benefits and costs is only partly convincing.

Of course within this class of problems there is the question of what mix of education is offered and chosen. Nelson and Wright report a marked decline in the US recently in the support for fundamental science. The same is true in the UK. Practically oriented people have never valued fundamentals – Ford pooh-poohed their pursuit and Mrs. Thatcher lamented that the U.K. produced too many Nobel Prizes. There are a number of conjectures of why this should be so, but there are also a number of R and D models in which variously trained labour appears in the functions, which argue what choices a growth maximising government should make. (AH and, for instance, Phelps [1966]).

It seems widely agreed that there are increasing returns in the acquisition of human capital (Lucas [1988]) and this allows for education to have an effect on growth[5]. Education is also meant to increase mobility of labour with the same consequence. But here other exogenously given (for a time) factors may work in the opposite direction as the distribution of the housing stock and preferences for locality. But one thing is clear: education is required for R and D, that is, it is now a necessary condition for growth. That was not true in the nineteenth century when R and D hardly existed, and yet there was much growth by learning by doing and shrewd management. So it depended on the (for us) exogenously given progress of science and engineering whether R and D is the main vehicle for growth by innovation. (See below.) Clearly there will be different mixes possible and to some extent the market will determine these if the educational manpower makes it feasible.

A much debated question is whether there are externalities to education. It is usual to suppose that more skill etc. is reflected in higher wages. Looking at the wages of engineers in the UK that is by no means obvious. (It clearly also depends on the number undertaking the particular training.) But if education contributes to the productivity of R and D, it is unlikely that its benefits will be fully internalised. Moreover Spence's [1973] famous work suggests that the amount and kind of education is unlikely to be optimal. There are also other simple externalities when an extra skilled worker raises the productivity of those already in place.

AH in their chapter on education use many common-sense ideas to put into their functional dress. They do not provide a precise theory of the development of the educational sector and indeed often discuss policy measures to improve this, which suggests that the endogenous market theory does not suffice. But it would be wrong to deny that the market has considerable influence. However when the amount of education in an economy changes it is likely that the decision to pursue education will also change independently of the pecuniary benefits. For instance, for many people higher education is a vehicle for social mobility and is also driven by a herd effect. These are all matters which make full endogenisation harder.

I have already mentioned Nelson and Wright's emphasis on organisation. Recognition of this led to business schools in the US and then in almost all European countries. Whether this can be explained as an equilibrium outcome I do not know. Once again there are surely herd effects. Moreover many firms prefer to do their own management training and rely on learning by doing. On the other hand most business schools are heavily subsidised by business, which argues in another direction. I find it hard to see how this development is to be fitted into an R and D theory, and certainly think imitation a plausible model – imitation of a spontaneous bright idea.

## 15.8 R and D again

I want to include a somewhat more formal problem which is too little discussed. It is my belief that just stating it will contribute to an understanding of endogenous growth theory.

Consider a firm $i$ and the orthodox theory of its production decision. We endow it in the first instance with a production set $Y_i$, a subset of commodity space. How do we interpret $Y_i$? Is it the set of activities known to $i$? Or should we straight away write $Y_i(t)$, the set known to $i$ at $t$? Clearly here we had better opt for the latter. But the commodity space of which $Y_i(t)$ is a subset should also be indexed by $t$ for obvious reasons. Note that by commodities I include labour of different types and efficiency. Or again we may write $\hat{Y}_i$ as its 'divine' set by which I mean the set of all activities that will ever be known to $i$, (embedded in a 'divine' commodity set). Then $Y_i(t) \subset \hat{Y}_i$.

Whichever more or less equivalent approach we adopt there remains a problem: how do we represent organisation (or management) choice. (See Arrow [1974] and Radner [1993].) We have learned a great deal in recent years of the reasons why production cannot be completely decentralised – bargaining costs, information and other transaction costs are cited. So we ought to include organisation in $Y$ (and $\hat{Y}$. If that involves costs, i.e. inputs, for instance of managers, then this will have to be taken into account. There are also intangibles, (e.g. organisation charts) quite apart from possible special managerial input. It is seen that the sheer act of description is not a simple one nor will the usual properties of production sets remain.

Now R and D is to provide information to $i$ of the divine set. (It may also be used strategically but matters are hard enough already.) The amount of R and D undertaken by $i$ will depend on expected costs and benefits. (One uses an arbitrary equation of AH.) I have already argued that these depend on the amount of related research done by others in the economy, as well as on what, for a better expression, one might call the 'research atmosphere' in the economy. Of course it also depends on more conventional inputs of specialised and ordinary labour, and on equipment. There are good arguments to suggest that R and D has an increasing returns segment in all of its inputs. But these increasing returns – or the way we look at returns – may refer only or largely to the probability of success.

But success in what? The endogenous macro-economic growth modellers lump it all into productivity. That however does not tell a firm in which direction to point its research. It seems to me that the past here has some explanatory power. Once there are computers, smaller and faster ones are natural projects. Before that it was not at all clear that research here was worthwhile. Once the genetic code was discovered the direction of further research and development was fairly obvious, at least in broad outline. There will also be many past failures to point to which one now seeks to avoid. As far as decision makers are concerned the past is given, as far as economists are concerned it can be painfully explained but it is a complex task. I do not think that enough attention has been given to the stochastic which we all agree is needed. We do not really know what is the process from which realisation of say fundamental new knowledge or clever new production ideas are drawn. As far as I can see we claim to have an endogenous (macro) theory where peculiarities 'wash out' and where the machine is driven by a search for profit. That is better than nothing, but since we have no idea what the functional forms are, it leaves a great deal to 'it fits the data so far'.

As far as benefit calculations are concerned one needs to specify the organisation of the firm and its risk attitudes. One also must hope that benefits are recognised and properly evaluated. At the moment R and D expenditure in the UK is falling and has fallen rather a lot already. Why? I suspect it has a good deal to do with the type of British manager, but there are almost surely more profound explanations. As far as I can see competition from abroad has increased because of the high pound and it may be that foreigners are regarded as so much more 'dynamic' that the obsolescence risk is put very high.

One of the difficulties with present theorising is that it makes it very difficult to explain why some executives are highly paid and others less so. They all form policy in the same way and indeed the latter could be put on a computer and all a firm then needs is to recognise the 'conditional' of a policy and press the right button. There is in other words far too little said about different abilities to evaluate probabilities of different people (and societies) and again different risk attitudes disappear. There is no Schumpeterian entrepreneur and the analysis is all routine maximisation. Of course I am not arguing that orthodox maximisation approach yields no useful insights – only that we need more before we claim to have an 'endogenous' theory.

Innovations, as Schumpeter argued, disrupt routine. Many people dislike such disruption. Comparing the US with the UK it seems clear that this dislike is less pronounced in the former. Why? Do we build that into our equations?

R and D activity maps into the probability space of the divine set. What that map is I do not think we know, even at the macro-level, and we choose maps which can either yield steady states or cycles or both. Since both have been observed this is not to be despised but it seems to me rather mad to do this over infinite time and with an unchanging stochastic process, let alone under perfect competition.

My conclusion is this: given a history and a market economy we can single out elements which are important in producing growth and that in a fairly orthodox fashion. To do more is fiction.

## 15.9 The global economy (an addendum)

Up to now I have had the traditional ‘economy’ of our textbooks in mind. Even in that case – supposing it to be closed to foreign trade, it will receive relevant technological signals from the outside world. But that is well understood. However if we are concerned with large geographical areas where trade is free and capital (and possibly labour) movements are free, then one needs to think again. Like other areas, the literature on the subject is well summarised by AH.

I am only concerned with what becomes of the exogenous–endogenous distinction. It would seem that the scope of the endogenous is enlarged simply because there will be a ‘washing out’ of the special occurrences and in particular it may be that the importance of risk attitudes is diminished. I do not really know. But the scope for cartelisation is also enlarged, which points in the opposite direction.

Clearly one is now discussing convergence. But quite often, because of increasing returns and externalities, one would expect divergence. Southern Italy is a good example and I fear that parts of Europe will have a similar fate relative to, say, Germany. But of course we shall no longer sensibly speak of countries provided labour mobility and capital mobility is large. Much here depends on local authorities and exogenously given elements like language.

I would hesitate to predict that the global economy can be modelled in a similar way to local economies, and I would find it hard to believe that the plausibility of functional forms, e.g. for education etc., will remain unaffected. On the other hand national characteristics, utility functions and organisation, may indeed converge and so, as I have said already, make an endogenous theory more possible. I have not seen any very formal discussion of this, but it is often mentioned. Note that this suggests for instance an endogenous theory of tastes and aspiration.

So I conclude by saying that for a global economy the ‘exogenous’, beyond obvious elements like climate, may be less important. However for a long time to come functional forms needed by an endogenous theory will have little but plausibility to support them.

## Acknowledgement

I am grateful to Kenneth Arrow and Robert Solow for comments.

## Notes

1 Kenneth Arrow (in conversation) interprets ‘endogenous’ in this way, that is, to get a growth equation on the basis of rational choice. He takes Solow’s given rate of growth as the point of departure. There is no monopoly in terminology and if ‘endogenous’ is taken to have the Arrow meaning this chapter has the wrong title. I however regard the word to apply to a theory which does not treat as exogenous what economic theory is in principle equipped to explain. Moreover one is likely to be led into error when the full set of factors taken to be exogenous is not enumerated.

2 If one postulates risk neutrality, stochastic elements present no technical problems.

3 Solow has pointed out to me that some innovations may be complementary to existing technology.

4 What is sometimes discussed is the relative efficiency of various organisation forms (largely connected with the absorption of information) but it is not integrated into what we are used to call the production set. Of course these findings suggest that there were important increasing returns to scale. But it may also be that a large market reduces risk.

5 This may be taken to mean that the productivity of educational resources increases with the general educational level of society or that the cost of acquiring extra education by an individual is a decreasing function of his educational level.

## References

Aghion, P. and Howitt, P. (1998) *Endogenous Growth Theory*, M.I.T. Press.

Arrow, K. J. (1962) 'The Economic Implications of Learning by Doing', *Review of Economic Studies*, 29, 155–73.

Arrow, K. J. (1974) *The Limits of Organisation*, Norton and Company.

Hahn, F. (1990) 'Solowian Growth Models' in *Growth Productivity Unemployment*, (ed.) by P. Diamond, MIT Press.

Jorgenson, D.W. (1961) 'Development of a Dual Economy', *Economic Journal*, 71, June, 309–34.

Kaldor, N. (1957) 'A Model of Economic Growth', *Economic Journal*, 57, 591–624.

Lucas, Robert. E. (1988) 'On the Mechanics of Economic Development', *Journal of Monetary Economics*, 22, 3–42.

Nelson and Wright, G. (1992) 'The Rise and Fall of American Technological Leadership', *Journal of Economic Literature*, 30.

Phelps, E. (1966) 'Investment in Humans, Technological Diffusion and Economic Growth', *American Economic Review*, 61, 69–75.

Radner, R. (1993) 'The Organisation of Decentralised Information Processing', *Econometrica*, 62, 1109–46.

Romer, P. M. (1986) 'Increasing Returns and Long-Run Growth', *Journal of Political Economy*, 94, 1002–37.

Romer, P. M. (1990) 'Endogenous Technological Change', *Journal of Political Economy*, 98, S71–S102

Solow, R. M. (1956) 'A Contribution to the Theory of Economic Growth', *Quarterly Journal of Economics*, 70, 65–94.

Spence, M. (1973) 'Job Market Signalling', *Quarterly Journal of Economics*, 87, 355–74.

# Index