Methods for Applied Macroeconomic Research

FABIO CANOVA

Contents

List of Figures

List of Tables

Preface

There has been a tremendous improvement over the last twenty years in the mathematical, statistical, probabilistic and computational tools available to applied macroeconomists. This extended set of tools has changed the way researchers have approached the problem of testing models, validate theories or simply collect regularities from the data. The rational expectation and the calibration revolutions have also forced researchers to try to build a more solid bridge between theoretical and applied work, a bridge which was often missing in much of the applied exercises conducted in the 1970s and the 1980s.

This books attempts to bring together dynamic general equilibrium theory, data analysis, advanced econometric and computational methods to provide a comprehensive set of techniques which can be used to address questions of interest to academics, business and central bank economists in the fields of macroeconomics, business cycle analysis, growth theory, monetary, financial, and international economics. The point of view taken is the one of an applied economist facing time series data (at times a panel of them, coming from different countries), who is interested in verifying the prediction of dynamic economic theories and in advising model builders and theorists on how to respecify existing constructions to obtain better match between the model and the data. The book illustrates a number of techniques which can be used to address the questions of interest, agnostically evaluates their usefulness in bringing out information relevant to the users, provides examples where the methods work (and where they don't) and points out problems when approaches developed for microeconomic data are used in time series frameworks.

Unavoidably, a modern treatment of such a complex topic requires a quantitative perspective, a solid dynamic theory background and the development of both empirical and numerical methods. A quantitative perspective is needed to give empirical content to theories; empirical methods must provide an effective link between economic theory and the data; numerical techniques help us to solve complicated dynamic stochastic general equilibrium (DSGE) models and to implement advanced econometric estimators, both in the classical and Bayesian tradition. In some cases empirical methods are intimately linked with the numerical procedure chosen to solve the model. In others, they are only constrained by the restrictions economic theory imposes on the data.

Given this background, the structure of this book is quite different from the typical graduate textbook both in macroeconomics and in econometrics. Rather than listing a series of estimators and their properties for different data generating processes, this book starts from a class of DSGE models, finds an approximate (linear) representation for the decision rules and describes methods needed to estimate/choose their parameters, to examine their fit to the data and to conduct interesting policy exercises. The first three chapters of the book are introductory and review material extensively used in later chapters. In particular, chapter 1 presents basic time series and probability concepts, a list of useful law of large numbers and central limit theorems, which are employed in the discussions of chapters 4 to 8, and gives a brief overview of the basic elements of spectral analysis, heavily used in chapters 3, 5 and 7. Chapter 2 presents a number of macroeconomic models currently

used in the profession and discusses numerical methods needed to solve them. Most of the examples and exercises of this book are based on versions of these models. Chapter 3 discusses procedures used to obtain interesting information about secular and cyclical fluctuations in the data.

In the remaining chapters we present various methodologies to confront models to the data and discuss how they can be used to address other interesting economic questions. Given our empirical perspective, formal results are often stated without proofs and emphasis is given to their use in particular macroeconomic applications. Chapter 4 describes minimalist vector autoregressive (VAR) approaches, where a limited amount of economic theory is used to structure the data. Chapter 5 presents limited information methodologies such as Generalized Methods of Moments (GMM), Simulated Method of Moments (SMM) and general simulation approaches. Chapter 6 examines full information Maximum Likelihood and in chapter 7 Calibration techniques are discussed. In chapter 8, we then branch into dynamic macro panel methods, which can be used to effectively study cross-country issues, and conclude the book with an extensive description of Bayesian methods and their use for VAR and panel VAR models, for advanced time series specifications, and for DSGE models (Chapters 9 to 11).

The approach of this book differs, for example, from the one of Hamilton (1994) or Hayashi (2002), both of which are primarily directed to econometricians and are not directly concerned with the question of validating dynamic economic models. The emphasis also differs from more macroeconomic oriented books like Sargent and Liungqvist (2001) or computationally oriented books like Judd (1998) or Miranda and Fackler (2002) in that empirical methods play a larger role and the connection between theory, numerical and empirical tools is explicitely spelled out.

The book is largely self-contained but presumes a basic knowledge of modern macroeconomic theory (say, one or two quarters of a Ph.D. course in macroeconomics), of standard econometrics (say, a quarter of a Ph. D. course in econometrics) and assumes that the reader has or will acquire in the process some programming skills (e.g., RATS, Matlab, Gauss). The book is thought for a year long sequence starting from second semester of a first year econometric/ applied macroeconomics course and continuing with the first semester of a second year macroeconometric course. Roughly, the first 5 chapters and the seventh could be thought in first part, chapter 6 and the last four in the second part. This is the setup I have used in teaching this material over a number years and it seems the natural division between what I consider basic and advanced material.

Ph. D. students at Brown University, University of Rochester, Universitat Pompeu Fabra, Universita' di Napoli, University of Porto, University of Southampton, London Business School, Bocconi University, Universita' Milano-Bicocca; participants in various editions of the Barcelona Summer School in Macroeconomics (BSSM), of the European Economic Association (EEA) Summer school in Macroeconomics, Paris, of the Center for Financial Studies (CFS) Summer school in Macroeconomics, Eltville (Germany), of the ZEI Summer School in Bonn, of the course for Central Bankers in Genzersee (Switzerland); and attendants of various intense and short courses at the ECB, Bank of England, Bank of Italy,

Bank of Canada, Bank of Hungary, Riksbank, Bundesbank and European Business Cycle Network (EABCN) have passed through several versions of this book and played around with some of the codes which implement the procedures discussed in the book with some practical examples. Some suffered; some enthusiastically embraced the philosophy of this book; some were critical; some made useful comments and helped in debugging the codes, all of them were encouraging. To all goes my thanks. I have learned a lot through the process of writing this book and teaching its material, probably as much as students have learned from the lectures and practical sessions.

Three people taught me to approach empirical problems in a sensible but rigorous way, combining economic theory with advanced statistical tools and numerical methods, and to be suspicious and critical of analyses which leave out one of the main ingredients of the cake. Christopher Sims and Tom Sargent where of crucial importance in making me understand that the grey area at the crossroad between theory and econometrics is a difficult but exciting place to be and their uncompromising intellectual curiosity, their stern intuition and their deep understanding of economic and policy issues has been an extraordinary lever behind this book. Adrian Pagan shaped my (somewhat cynical) view of what should and can be done with the data and the models. I always like to argue with him because his unconventional views helped to bring out often forgotten methodological and practical aspects. And on most issues of interest to applied macroeconomists he was more often right than wrong. This book would not have been possible without their fundamental inputs. As mentors, there was no one comparable to them. I also have an intellectual debit with Ed Prescott. It was his brusc refusal to follow the traditional econometric track that made me understand the need to create a different and more solid link between theory, econometric and statistical techniques and the data. Several of my collegues, in particular Albert Marcet and Morten Ravn, Jordi Gali, Lucrezia Reichlin, Harald Uhlig, Carlo Favero, Marco Maffezzoli and Luca sala contributed to form and develop some of the ideas presented in the book. A special thanks goes to Tom Doan, Marco del Negro, Chris Sims, Kirdan Lees and Adrian Pagan, who spotted mistakes and imprecisions in earlier versions of the manuscript.

Writing a textbook is difficult. Writing an advanced textbook, which bring together material from different fields, is even more formidable. Many times I run out of steam, I got bored and was ready to give up and do something different. Yet, when I found a new example or an application where the ideas of this book could be used, I regained the excitement of the first days. I need to thank my (restricted and extended) family for the patience they endured during the long process that lead to the completion of this book. Dynamic macroeconomics is in part about intertemporal substitution. Patience is probably built on the same principle.

Chapter 1: Preliminaries

This chapter is introductory and it is intended for readers who are unfamiliar with time series concepts, with the properties of stochastic processes, with basic asymptotic theory results and with the a-b-c of spectral analysis. Those who feel comfortable with these topics can skip it and directly proceed to chapter 2.

Since the material is vast and complex, an effort is made to present it at the simplest possible level, emphasizing a selected number of topics and only those aspects which are useful for the central topic of this book: comparing the properties of dynamic stochastic general equilibrium (DSGE) models to the data. This means that intuition rather than mathematical rigor is stressed. More specialized books, such as Brockwell and Davis (1990), Davidson (1994), Priestley (1980) or White (1984), provide a comprehensive and in-depth treatment of these topics.

When trying to provide background material, there is always the risk of going too far back to the basics and so to speak "attempt to reinvent the wheel". To avoid this, we assume that the reader knows simple calculus concepts like limits, continuity and uniform continuity of functions of real numbers and that she is familiar with distributions functions and measures.

The chapter is divided in six sections. The first defines what a stochastic process is. The second examines the limiting behavior of stochastic processes introducing four concepts of convergence and characterizing their relationships. Section 3 deals with time series concepts. Section 4 deals with laws of large numbers. These laws are useful to insure that functions of stochastic processes converge to appropriate limits. We examine three situations: a case where the elements of a stochastic process are dependent and identically distributed; one where they are dependent and heterogeneously distributed and one where they are martingale differences. Section 5 describes three central limit theorems corresponding to the three situations analyzed in section 4. Central limit theorems are useful to derive the limiting distribution of functions of stochastic processes and are the basis for (classical) tests of hypotheses and for some model evaluation criteria.

Section 6 presents elements of spectral analysis. Spectral analysis is useful for breaking down economic time series into components (trends, cycles, etc.), for building measures of persistence in response to shocks, for computing the asymptotic covariance matrix of certain estimators and for defining measures of distance between a model and the data. It may be challenging at first. However, once it is realized that most of the functions typically

performed in everyday life employ spectral methods (frequency modulation in a stereo; frequency band reception in a cellular phone, etc.), the reader should feel more comfortable with it. Spectral analysis offers an alternative way to look at time series, translating serially dependent time observations in contemporaneously independent frequency observations. This change of coordinates allows us to analyze the primitive cycles which compose time series, discuss their length, amplitude and persistence.

Whenever not explicitly stated, the machinery presented in this chapter applies to both scalar and vector stochastic processes. The notation $\{y_t(\varkappa)\}_{t=-\infty}^{\infty}$ indicates the sequence $\{\ldots, y_0(\varkappa), y_1(\varkappa), \ldots, y_t(\varkappa), \ldots\}$ where, for each t, the random variable $y_t(\varkappa)$ is a measurable function [1] of the state of nature $\varkappa$, i.e. $y_t(\varkappa) : \mathbb{K} \to \mathbb{R}$, where $\mathbb{R}$ is the real line and $\mathbb{K}$ the space of states of nature. We also assume that at each τ $\{y_\tau(\varkappa)\}_{\tau=-\infty}^{t}$ is known so that any function $h(y_\tau)$ will be "adapted" to this information structure. To simplify the notation at times we write $\{y_t(\varkappa)\}$ or y_t. A normal random variable with zero mean and variance Σ_y is denoted by $y_t \sim \mathbb{N}(0, \Sigma_y)$ and a random variable uniformly distributed over the interval $[a_1, a_2]$ is denoted by $y_t \sim \mathbb{U}[a_1, a_2]$. Finally, *iid* indicates identically and independently distributed random variables.

1.1 Stochastic Processes

Definition 1.1 *(Stochastic Process): A stochastic process* $\{y_t(\varkappa)\}_{t=1}^{\infty}$ *is a probability measure defined on sets of sequences of real vectors (the "paths" of the process).*

The definition implies, among other things, that the set of paths $\mathbb{X} = \{y : y_t(\varkappa) \leq \varrho\}$, for arbitrary $\varrho \in \mathbb{R}$, t fixed, has well-defined probabilities. In other words, choosing different $\varrho \in \mathbb{R}$ for a given t, and performing countable unions, finite intersections and complementing the above set of paths, we generate a set of events with proper probabilities. Note that y_t is unrestricted for all $\tau \leq t$: the realization needs not to exceed ϱ only at t. Observable time series are realizations of a stochastic process $\{y_t(\varkappa)\}_{t=1}^{\infty}$, given $\varkappa$ [2].

Example 1.1 *Two simple stochastic processes are the following:*
1) $\{y_t(\varkappa)\} = e_1 \cos(t \times e_2)$*, where* e_1, e_2 *are random variables,* $e_1 > 0$ *and* $e_2 \sim \mathbb{U}[0, 2\pi)$*,* $t > 0$*. Here* y_t *is periodic:* e_1 *controls the amplitude and* e_2 *the periodicity of* y_t*.*
2) $\{y_t(\varkappa)\}$ *is such that* $P[y_t = \pm 1] = 0.5$ *for all* t*. Such a process has no memory and flips between -1 and 1 as* t *changes.*

Example 1.2 *It is easy to generate complex stochastic processes from primitive ones. For example, if* $e_{1t} \sim \mathbb{N}(0, 1)$*,* $e_{2t} \sim \mathbb{U}(0, 1]$ *and independent of each other,* $y_t = e_{2t} \exp\{\frac{e_{1t}}{1+e_{1t}}\}$ *is a stochastic process. Similarly,* $y_t = \sum_{t=1}^{T} e_t$*,* e_t *iid* $\sim (0, 1)$ *is a stochastic process.*

---

[1] A function of h is $\mathcal{F}$-measurable if for every real number ϱ, the set $[\varkappa : h(\varkappa) < \varrho]$ belongs to $\mathcal{F}$, where $\mathcal{F}$ is typically chosen to be the collection of Borel sets of $\mathbb{R}_\infty = \mathbb{R} \times \mathbb{R} \times \ldots$.

[2] A stochastic process could also be defined as a sequence of random variables which are jointly measurable, see e.g. Davidson, 1994, p. 177.

1.2 Concepts of Convergence

In a classical framework the properties of estimators are obtained using sequences of estimators indexed by the sample size, and showing that these sequences approach the true (unknown) parameter value as the sample size grows to infinity. Since estimators are continuous functions of the data, we need to insure that the data possesses a proper limit and that continuous functions of the data inherit these properties. To show that the former is the case one can rely on a variety of convergence concepts. The first two deal with convergence of the sequence, the next with its moments and the latter with its distribution.

1.2.1 Almost sure (a.s.) convergence

The concept of a.s. convergence extends the idea of convergence to a limit employed in the case of a sequence of real numbers.

As we have seen, the elements of the sequence $y_t(\varkappa)$ are functions of the state of nature. However, once $\varkappa = \bar{\varkappa}$ is drawn, $\{y_1(\bar{\varkappa}), \ldots, y_t(\bar{\varkappa}), \ldots\}$ looks like a sequence of real numbers. Hence, given $\varkappa = \bar{\varkappa}$, convergence can be similarly defined.

Definition 1.2 *(almost sure convergence)* $y_t(\varkappa)$ *converges almost surely to* $y(\varkappa) < \infty$, *denoted by* $\{y_t(\varkappa)\} \overset{a.s.}{\to} y(\varkappa)$, *if* $\lim_{T\to\infty} P[||y_t(\varkappa) - y(\varkappa)|| \leq \varepsilon, \ \forall t > T] = 1$, *for* $\varkappa \in \mathbb{K}_1 \subseteq \mathbb{K}$, *and every* $\varepsilon > 0$.

According to definition 1.2 $\{y_t(\varkappa)\}$ converges almost surely (a.s.) if the probability of obtaining a path for y_t which converges to $y(\varkappa)$ after some T is one. The probability is taken over $\varkappa$'s. The definition implies that failure to converge is possible, but it will almost never happen. When $\mathbb{K}$ is infinite dimensional, a.s. convergence is called convergence almost everywhere; sometimes a.s. convergence is termed convergence with probability 1 or strong consistency criteria.

Next, we describe the limiting behavior of functions of a.s. convergent sequences.

Result 1.1 *Let* $\{y_t(\varkappa)\} \overset{a.s.}{\to} y(\varkappa)$. *Let* h *be a* $n \times 1$ *vector of functions, continuous at* $y(\varkappa)$. *Then* $h(y_t(\varkappa)) \overset{a.s.}{\to} h(y(\varkappa))$. □

Result 1.1 is a simple extension of the standard fact that continuous functions of convergent sequences are convergent.

Example 1.3 *Given* $\varkappa$, *let* $\{y_t(\varkappa)\} = 1 - \frac{1}{t}$ *and let* $h(y_t(\varkappa)) = \frac{1}{T}\sum_t y_t(\varkappa)$. *Then* $h(y_t(\varkappa))$ *is continuous at* $\lim_{t\to\infty} y_t(\varkappa) = 1$ *and* $h(y_t(\varkappa)) \overset{a.s.}{\to} 1$.

Exercise 1.1 *Suppose* $\{y_t(\varkappa)\} = 1/t$ *with probability* $1 - 1/t$ *and* $\{y_t(\varkappa)\} = t$ *with probability* $1/t$. *Does* $\{y_t(\varkappa)\}$ *converge almost surely to 1?*

In some applications we will be interested in examining situations where a.s. convergence does not hold. This can be the case when the observations have a probability density

function that changes over time or when matrices appearing in the formula for estimators do not converge to fixed limits. However, even though $h(y_{1t}(\varkappa))$ does not converge to $h(y(\varkappa))$, it may be the case that the distance between $h(y_{1t}(\varkappa))$ and $h(y_{2t}(\varkappa))$ becomes arbitrarily small as $t \to \infty$, where $\{y_{2t}(\varkappa)\}$ is another sequence of random variables. To obtain convergence in this situation we need to strengthen the conditions by requiring uniform continuity of h (for example, assuming continuity on a compact set).

Result 1.2 *Let h be continuous on a compact set $\mathbb{R}_2 \in \mathbb{R}^m$. Suppose that $\{y_{1t}(\varkappa)\} - \{y_{2t}(\varkappa)\} \overset{a.s.}{\to} 0$ and there exists an $\epsilon > 0$ such that for all $t > T$, $[|y_{1t} - y_{2t}| < \epsilon] \subset \mathbb{R}_2$, t large. Then $h(y_{1t}(\varkappa)) - h(y_{2t}(\varkappa)) \overset{a.s.}{\to} 0$.* □

One application of result 1.2 is the following: suppose $\{y_{1t}(\varkappa)\}$ is some actual time series and $\{y_{2t}(\varkappa)\}$ is its counterpart simulated from a model where the parameters of the model and $\varkappa$ are given, and let h be some continuous statistics, e.g. the mean or the variance. Then, result 1.2 tells us that if simulated and actual paths are close enough as $t \to \infty$, statistics generated from these paths will also be close.

1.2.2 Convergence in Probability

Convergence in probability is weaker concept than almost sure convergence.

Definition 1.3 *(Convergence in Probability) If there exists a $y(\varkappa) < \infty$ such that, for every $\epsilon > 0$, $P[\varkappa : ||y_t(\varkappa) - y(\varkappa)|| < \epsilon] \to 1$, for $t \to \infty$, then $\{y_t(\varkappa)\} \overset{P}{\to} y(\varkappa)$.*

$\overset{P}{\to}$ is weaker than $\overset{a.s.}{\to}$ because in the former we only need the joint distribution of $(y_t(\varkappa), y(\varkappa))$ not the joint distribution of $(y_t(\varkappa), y_\tau(\varkappa), y(\varkappa))$ $\forall \tau > T$. $\overset{P}{\to}$ implies that it is less likely that one element of the $\{y_t(\varkappa)\}$ sequence is more than an ϵ away from $y(\varkappa)$ as $t \to \infty$. $\overset{a.s.}{\to}$ implies that after T, the path of $\{y_t(\varkappa)\}$ is not far from $y(\varkappa)$ as $T \to \infty$. Hence, it is easy to build examples where $\overset{P}{\to}$ does not imply $\overset{a.s.}{\to}$.

Example 1.4 *Let y_t and y_τ be independent $\forall t, \tau$, let y_t be either 0 or 1 and let*

$$P[y_t = 0] = \begin{cases} 1/2 & t = 1, 2 \\ 2/3 & t = 3, 4 \\ 3/4 & t = 5 \cdots 8 \\ 4/5 & t = 9 \cdots 16 \end{cases}$$

Then $P[y_t = 0] = 1 - \frac{1}{j}$ for $t = 2^{(j-1)+1}, \ldots, 2^j$, $j > 1$, so that $y_t \overset{P}{\to} 0$. This is because the probability that y_t is in one of these classes is $1/j$ and, as $t \to \infty$, the number of classes goes to infinity. However, y_t does not converge almost surely to zero since the probability that a convergent path is drawn is zero; i.e., if at t we draw $y_t = 1$, there is a non-negligible probability that $y_{t+1} = 1$ is drawn. In general, $y_t \overset{P}{\to} 0$ is too slow to insure that $y_t \overset{a.s.}{\to} 0$.

Although convergence in probability does not imply almost sure convergence, the following result shows how the latter can be obtained from the former.

Result 1.3 *If* $y_t(\varkappa) \xrightarrow{P} y(\varkappa)$, *there exist a subsequence* $y_{t_j}(\varkappa)$ *such that* $y_{t_j}(\varkappa) \xrightarrow{a.s.} y(\varkappa)$ *(see, e.g., Lukacs, 1975, p. 48).* □

Intuitively, since convergence in probability allows a more erratic behavior in the converging sequence than almost sure convergence, one can obtain the latter by disregarding the erratic elements. The concept of convergence in probability is useful to show "weak" consistency of certain estimators.

Example 1.5 *(i) Let* y_t *be a sequence of iid random variables with* $E(y_t) < \infty$. *Then* $\frac{1}{T}\sum_{t=1}^{T} y_t \xrightarrow{a.s.} E(y_t)$ *(Kolmogorov strong law of large numbers).*

(ii) Let y_t *be a sequence of uncorrelated random variables,* $E(y_t) < \infty$, $var(y_t) = \sigma_y^2 < \infty$, $cov(y_t, y_{t-\tau}) = 0$ *for all* $\tau \neq 0$. *Then* $\frac{1}{T}\sum_{t=1}^{T} y_t \xrightarrow{P} E(y_t)$ *(Chebyshev weak law of large numbers).*

In example 1.5 strong consistency requires iid random variables, while for weak consistency we just need a set of uncorrelated random variables with identical means and variances. Note also that, weak consistency requires restrictions on the second moments of the sequence which are not needed in the former case.

The analogs of results 1.1 and 1.2 for convergence in probability can be easily obtained.

Result 1.4 *Let* $\{y_t(\varkappa)\}$ *be such that* $\{y_t(\varkappa)\} \xrightarrow{P} y(\varkappa)$. *Then* $h(y_t(\varkappa)) \xrightarrow{P} h(y(\varkappa))$, *for any* h *continuous at* $y(\varkappa)$ *(see White, 1984, p. 23).* □

Result 1.5 *Let* h *be continuous on a compact* $\mathbb{R}_2 \subset \mathbb{R}^m$. *Let* $\{y_{1t}(\varkappa)\}$ *and* $\{y_{2t}(\varkappa)\}$ *be such that* $\{y_{1t}(\varkappa)\} - \{y_{2t}(\varkappa)\} \xrightarrow{P} 0$ *for* $t \to \infty$. *Then* $h(y_{1t}(\varkappa)) - h(y_{2t}(\varkappa)) \xrightarrow{P} 0$ *(see White, 1984, p. 25).* □

Sometimes y_t may converge to a limit which does not belong to the space of the random variables which make up the sequence; e.g., the sequence $y_t = \sum_j e_j$, where each e_j is iid has a limit which is not in the space of iid variables. In other cases, the limit point may be unknown. For all these cases, we can redefine almost sure convergence and convergence in probability using the concept of Cauchy sequences.

Definition 1.4 *(Convergence a.s and in Probability):* $\{y_t(\varkappa)\}$ *converges a.s. if and only if for every* $\epsilon > 0$, $\lim_{T\to\infty} P[||y_t(\varkappa) - y_\tau(\varkappa)|| > \epsilon$, *for some* $\tau > t \geq T(\varkappa, \epsilon)] \to 0$ *and converges in probability if and only if for every* $\epsilon > 0$, $\lim_{t,\tau\to\infty} P[||y_t(\varkappa) - y_\tau(\varkappa)|| > \epsilon] \to 0$.

1.2.3 Convergence in L^q-norm.

While almost sure convergence and convergence in probability concern the path of y_t, L^q-convergence refers to the q-th moment of y_t. L^q-convergence is typically analyzed when $q = 2$ (convergence in mean square); when $q = 1$ (absolute convergence); and when $q = \infty$ (minmax convergence).

Definition 1.5 *(Convergence in the norm): $\{y_t(\varkappa)\}$ converges in the L^q-norm, (or, in the q^{th}-mean), denoted by $y_t(\varkappa) \overset{q.m.}{\to} y(\varkappa)$, if there exists a $y(\varkappa) < \infty$ such that $\lim_{t\to\infty} E[|y_t(\varkappa) - y(\varkappa)|^q] = 0$, for some $q > 0$.*

Obviously, if the qth-moment does not exist, convergence in L^q does not apply (i.e., if y_t is a Cauchy random variable, L^q convergence is meaningless for all q), while convergence in probability applies even when moments do not exist. Intuitively, the difference between the two types of convergence lies in the fact that the latter allows the distance between y_t and y to get large faster than the probability gets smaller, while this is not possible with L^q convergence.

Exercise 1.2 *Let y_t converge to 0 in L^q. Show that y_t converges to 0 in probability. (Hint: Use Chebyshev's inequality.)*

Exercise 1.2 indicates that L^q convergence is stronger than convergence in probability.

The following result provides conditions insuring that convergence in probability imply L^q-convergence.

Result 1.6 *If $y_t(\varkappa) \overset{P}{\to} y(\varkappa)$ and $\sup_t\{\lim_{\Delta\to\infty} E(|y_t|^q \mathcal{I}_{[|y_t|\geq\Delta]})\} = 0$, where $\mathcal{I}$ is an indicator function, then $y_t(\varkappa) \overset{q.m.}{\to} y(\varkappa)$ (Davidson, 1994, p.287).* □

Hence, convergence in probability plus the restriction that $|y_t|^q$ is uniformly integrable, insures convergence in the L^q-norm. In general, there is no relationship between L^q and almost sure convergence. The following shows that the two concepts are distinct.

Example 1.6 *Let $y_t(\varkappa) = t$ if $\varkappa \in [0, 1/t)$ and $y_t(\varkappa) = 0$ otherwise. Then the set $\{\varkappa : \lim_{t\to\infty} y_t(\varkappa) \neq 0\}$ includes only the element $\{0\}$ so $y_t \overset{a.s.}{\to} 0$. However $E|y_t|^q = 0 * (1 - 1/t) + t^q/t = t^{q-1}$. Since y_t is not uniformly integrable it fails to converge in the q-mean for any $q > 1$ (for $q = 1, E|y_t| = 1, \forall t$). Hence, the limiting expectation of y_t differs from its almost sure limit.*

Exercise 1.3 *Let*

$$y_t = \begin{cases} 1 & \text{with probability } 1 - 1/t^2 \\ t & \text{with probability } 1/t^2 \end{cases}$$

Show that the first and second moments of y_t are finite. Show that $y_t \overset{P}{\to} 1$ but that y_t does not converge in quadratic mean to 1.

The next result shows that convergence in the $L^{q'}$-norm obtains when we know that convergence in the L^q-norm occurs, $q > q'$. The result makes use of Jensen's inequality, which we state next: Let h be a convex function on $\mathbb{R}_1 \subset \mathbb{R}^m$ and y be a random variable such that $P[y \in \mathbb{R}_1] = 1$. Then $h[E(y)] \leq E(h(y))$. If h is concave on $\mathbb{R}_1$, $h(E(y)) \geq E(h(y))$.

Example 1.7 *For* $h(y) = y^{-2}$, $Eh(y) = E(y^{-2}) \leq 1/E(y^2) = h(E(y))$.

Result 1.7 *Let* $q' < q$. *If* $y_t(\varkappa) \overset{q.m.}{\to} y(\varkappa)$, *then* $y_t(\varkappa) \overset{q'.m.}{\to} y(\varkappa)$. □

Example 1.8 *Let* $\mathbb{K} = \{\varkappa_1, \varkappa_2\}$ *and* $P(\varkappa_1) = P(\varkappa_2) = 0.5$. *Let* $y_t(\varkappa_1) = (-1)^t$, $y_t(\varkappa_2) = (-1)^{t+1}$ *and let* $y(\varkappa_1) = y(\varkappa_2) = 0$. *Clearly,* y_t *converges in the* L^q*-norm. To confirm this note, for example, that* $\lim_{t\to\infty} E[|y_t(\varkappa) - y(\varkappa)|^2] = 1$. *Since* y_t *converges in mean square, it must converge in absolute mean. In fact,* $\lim_{t\to\infty} E[|y_t(\varkappa) - y(\varkappa)|] = 1$.

1.2.4 Convergence in Distribution

Definition 1.6 *(Convergence in Distribution): Let* $\{y_t(\varkappa)\}$ *be a* $m \times 1$ *vector with joint distribution* $\mathcal{D}_t$. *If* $\mathcal{D}_t(z) \to \mathcal{D}(z)$ *as* $t \to \infty$, *for every point of continuity* z, *where* $\mathcal{D}$ *is the distribution function of a random variable* $y(\varkappa)$, *then* $y_t(\varkappa) \overset{D}{\to} y(\varkappa)$.

Convergence in distribution is the weakest convergence concept and does not imply, in general, anything about the convergence of a sequence of random variables. Moreover, while the previous three convergence concepts require $\{y_t(\varkappa)\}$ and the limit $y(\varkappa)$ to be defined on the same probability space, convergence in distribution is meaningful even when this is not the case.

It is useful to characterize the relationship between convergence in distribution and convergence in probability.

Result 1.8 *Suppose* $y_t(\varkappa) \overset{P}{\to} y(\varkappa) < \infty$, $y(\varkappa)$ *constant. Then* $y_t(\varkappa) \overset{D}{\to} \mathcal{D}_y$, *where* $\mathcal{D}_y$ *is the distribution of a random variable* z *such that* $P[z = y(\varkappa)] = 1$. *Conversely, if* $y_t(\varkappa) \overset{D}{\to} \mathcal{D}_y$, *then* $y_t \overset{P}{\to} y$ *(see Rao, 1973, p. 120).* □

Note that the first part of result **??** could have been obtained directly from result 1.4, had we assumed that $\mathcal{D}_y$ is a continuous function of y.

The next two results are handy when demonstrating the limiting properties of a class of estimators in dynamic models. Note that $y_{1t}(\varkappa)$ is $O_p(t^j)$ if there exists an $O(1)$ nonstochastic sequence y_{2t} such that $(\frac{1}{t^j} y_{1t}(\varkappa) - y_{2t}) \overset{P}{\to} 0$ and that y_{2t} is $O(1)$ if for some $0 < \Delta < \infty$, there exists a T such that $|y_{2t}| < \Delta$ for all $t \geq T$.

Result 1.9 *(i) If* $y_{1t} \xrightarrow{P} \varrho$, $y_{2t} \xrightarrow{D} y$ *then* $y_{1t}y_{2t} \xrightarrow{D} \varrho y$, $y_{1t} + y_{2t} \xrightarrow{D} \varrho + y$, *where* ϱ *is a constant (Davidson, 1994, p.355).*

(ii) If y_{1t} *and* y_{2t} *are sequences of random vectors,* $y_{1t} - y_{2t} \xrightarrow{P} 0$ *and* $y_{2t} \xrightarrow{D} y$ *imply that* $y_{1t} \xrightarrow{D} y$. *(Rao, 1973, p.123)* □

Part (ii) of result 1.9 is useful when the distribution of y_{1t} cannot be determined directly. In fact, if we can find a y_{2t} with known asymptotic distribution, which converges in probability to y_{1t}, then the distributions of y_{1t} and y_{2t} will coincide. We will use this result in chapter 5 when discussing two-steps estimators.

The limiting behavior of continuous functions of sequences which converge in distribution is easy to characterize. In fact we have:

Result 1.10 *Let* $y_t \xrightarrow{D} y$. *If* h *is continuous,* $h(y_t) \xrightarrow{D} h(y)$ *(Davidson, 1994, p. 355).* □

1.3 Time Series Concepts

Most of the analysis conducted in this book assumes that observable time series are stationary and have memory which dies out sufficiently fast over time. In some cases we will use alternative and weaker hypotheses which allow for selected forms of non-stationarity and/or for more general memory requirements. This section provides definitions of these concepts and compare various alternatives.

We need two preliminary definitions:

Definition 1.7 *(Lag operator): The lag operator is defined by* $\ell y_t = y_{t-1}$ *and* $\ell^{-1}y_t = y_{t+1}$. *When applied to a sequence of* $m \times m$ *matrices* $A_j, j = 1, 2, \ldots$, *the lag operator produces* $A(\ell) = A_0 + A_1\ell + A_2\ell^2 + \ldots$.

Definition 1.8 *(Autocovariance function): The autocovariance function of* $\{y_t(\varkappa)\}_{t=-\infty}^{\infty}$ *is* $ACF_t(\tau) \equiv E(y_t(\varkappa) - E(y_t(\varkappa)))(y_{t-\tau}(\varkappa) - E(y_{t-\tau}(\varkappa)))$ *and its autocorrelation function* $ACRF_t(\tau) \equiv \text{corr}(y_t, y_{t-\tau}) = \frac{ACF_t(\tau)}{\sqrt{var(y_t(\varkappa))var(y_{t-\tau}(\varkappa))}}$.

In general, both the autocovariance and the autocorrelation functions depend on time and on the gap between y_t and $y_{t-\tau}$.

Definition 1.9 *(Stationarity 1):* $\{y_t(\varkappa)\}_{t=-\infty}^{\infty}$ *is stationary if for any set of paths* $\mathbb{X} = \{y_t(\varkappa) : y_t(\varkappa) \leq \varrho, \varrho \in \mathbb{R}, \varkappa \in \mathbb{K}\}$, $P(\mathbb{X}) = P(\ell^\tau\mathbb{X})$, $\forall\tau$, *where* $\ell^\tau\mathbb{X} = \{y_{t-\tau}(\varkappa) : y_{t-\tau}(\varkappa) \leq \varrho\}$.

A process is stationary if shifting a path over time does not change the probability distribution of that path. In this case the joint distribution of $\{y_{t_1}, \ldots, y_{t_j}\}$ is the same as the joint distribution of $\{y_{t_{1+\tau}}, \ldots, y_{t_{j+\tau}}\}$, $\forall\tau$. A weaker concept is the following:

Definition 1.10 *(Stationarity 2): $\{y_t(\varkappa)\}_{t=-\infty}^{\infty}$ is covariance (weakly) stationary if $E(y_t)$ is constant; $E|y_t|^2 < \infty$; $ACF_t(\tau)$ is independent of t.*

Definition 1.10 is weaker than 1.9 in several senses: first, it involves the distribution of y_t at each t and not the joint distribution of a (sub)sequence of y_t's. Second, it only concerns the first two moments of y_t. Clearly, a stationary process is weakly stationary, while the converse is true, only when y_t's are normal random variables. In fact, when y_t is normal, the first two moments characterize the entire distribution and the joint distribution of a $\{y_t\}_{t=1}^{\infty}$ path is normal.

Example 1.9 *Let $y_t = e_1\cos(\omega t)+e_2\sin(\omega t)$, where e_1, e_2 are uncorrelated with mean zero, unit variance and $\omega \in [0, 2\pi]$. Clearly, the mean of y_t is constant and $E|y_t|^2 < \infty$. Also $cov(y_t, y_{t+\tau}) = \cos(\omega t)\cos(\omega(t+\tau))+\sin(\omega t)\sin(\omega(t+\tau)) = \cos(\omega\tau)$. Hence y_t is covariance stationary.*

Exercise 1.4 *Suppose $y_t = e_t$ if t is odd and $y_t = e_t + 1$ if t is even, where $e_t \sim$ iid $(0, 1)$. Show that y_t is not covariance stationary. Show that $y_t = \bar{y} + y_{t-1} + e_t$, $e_t \sim$ iid $(0, \sigma_e^2)$, where $\bar{y}$ is a constant is not stationary but that $\Delta y_t = y_t - y_{t-1}$ is stationary.*

When $\{y_t\}$ is stationary, its autocovariance function has three properties: (i) $ACF(0) \geq 0$, (ii) $|ACF(\tau)| \leq ACF(0)$, (iii) $ACF(-\tau) = ACF(\tau)$ for all τ. Furthermore, if y_{1t} and y_{2t} are two stationary uncorrelated sequences, $y_{1t}+y_{2t}$ is stationary and the autocovariance function of $y_{1t} + y_{2t}$ is $ACF_{y_1}(\tau) + ACF_{y_2}(\tau)$.

Example 1.10 *Consider the process $y_t = \bar{y} + at + De_t$, where $|D| < 1$ and $e_t \sim (0, \sigma^2)$. Clearly y_t is not covariance stationary since $E(y_t) = \bar{y}+at$, which depends on time. Taking first difference we have $\Delta y_t = a + D\Delta e_t$. Here $E(\Delta y_t) = a$, $E(\Delta y_t - a)^2 = 2D^2\sigma^2 > 0$, $E(\Delta y_t - a)(\Delta y_{t-1} - a) = -D^2\sigma^2 < E(\Delta y_t - a)^2$, and $E(\Delta y_t - a)(\Delta y_{t+1} - a) = -D^2\sigma^2$.*

Exercise 1.5 *Suppose $y_{1t} = \bar{y} + at + e_t$, where $e_t \sim$ iid $(0, \sigma_e^2)$ and $\bar{y}, a$ are constants. Define $y_{2t} = \frac{1}{2J+1}\sum_{j=-J}^{J} y_{1t+j}$. Compute the mean and the autocovariance function of y_{2t}. Is y_{2t} stationary? Is it covariance stationary?*

Definition 1.11 *(Autocovariance generating function): The autocovariance generating function of a stationary $\{y_t(\varkappa)\}_{t=-\infty}^{\infty}$ is $CGF(z) = \sum_{\tau=-\infty}^{\infty} ACF(\tau)z^{\tau}$, provided that the sum converges for all z satisfying $\varrho^{-1} < |z| < \varrho$, $\varrho > 1$.*

Example 1.11 *Consider the process $y_t = e_t - De_{t-1} = (1 - D\ell)e_t$, $|D| < 1$, $e_t \sim$ iid $(0, \sigma_e^2)$. Here $cov(y_t, y_{t-j}) = cov(y_t, y_{t+j}) = 0, \forall j \geq 2$; $cov(y_t, y_t) = (1 + D^2)\sigma_e^2$; $cov(y_t, y_{t-1}) = -D\ell\sigma_e^2$; $cov(y_t, y_{t+1}) = -D\ell^{-1}\sigma_e^2$. Hence*

$$\begin{aligned} CGF_y(z) &= -D\sigma_e^2 z^{-1} + (1 + D^2)\sigma_e^2 z^0 - D\sigma_e^2 z^1 \\ &= \sigma_e^2(-Dz^{-1} + (1 + D^2) - Dz) = \sigma_e^2(1 - Dz)(1 - Dz^{-1}) \end{aligned} \tag{1.1}$$

Example 1.11 can be generalized to more complex processes. In fact, if $y_t = D(\ell)e_t$, $CFG_y(z) = D(z)\Sigma_e D(z^{-1})'$, and this holds for both univariate and multivariate y_t. One interesting special case occurs when $z = e^{-i\omega} = cos(\omega) - i\ sin(\omega),\ \omega \in (0, 2\pi),\ i = \sqrt{-1}$, in which case $\mathcal{S}(\omega) \equiv \frac{GCF_y(e^{-i\omega})}{2\pi} = \frac{1}{2\pi}\sum_{\tau=-\infty}^{\infty} ACF(\tau)e^{-i\omega\tau}$ is the spectral density of y_t.

Exercise 1.6 *Consider $y_t = (1 + 0.5\ell + 0.8\ell^2)e_t$, and $(1 - 0.25\ell)y_t = e_t$ where $e_t \sim$ iid $(0, \sigma_e^2)$. Are these processes covariance stationary? If so, show the autocovariance and the autocovariance generating functions.*

Exercise 1.7 *Let $\{y_{1t}(\varkappa)\}$ be a stationary process and let h be a $n \times 1$ vector of continuous functions. Show that $y_{2t} = h(y_{1t})$ is also stationary.*

Stationarity is a weaker requirement than iid, where no dependence between elements of a sequence is allowed, but it is stronger that the identically (not necessarily independently) distributed assumption.

Example 1.12 *Let $y_t \sim$ iid $(0, 1)\ \forall t$. Since $y_{t-\tau} \sim$ iid $(0, 1), \forall\tau$ any finite subsequence $y_{t_{1+\tau}}, \ldots, y_{t_{j+\tau}}$ will have the same distribution and therefore y_t is stationary. It is easy to see that a stationary series is not necessarily iid. For instance, let $y_t = e_t - De_{t-1}$. If $|D| < 1$, y_t is stationary but not iid.*

Exercise 1.8 *Give an example of a process y_t which is identically (but not necessarily independently) distributed which is nonstationary.*

A property of stationary sequences which insures that the sample average converges to the population average is ergodicity. Ergodicity is typically defined in terms of invariant events.

Definition 1.12 *(Ergodicity 1): Suppose $y_t(\varkappa) = y_1(\ell^{t-1}\varkappa)$, all t. Then $\{y_t(\varkappa)\}$ is ergodic if and only if for any set of paths $\mathbb{X} = \{y_t(\varkappa) : y_t(\varkappa) \leq \varrho, \varrho \in \mathbb{R}\}$, with $P(\ell^\tau\mathbb{X}) = P(\mathbb{X}),\ \forall\tau$, $P(\mathbb{X}) = 0$ or $P(\mathbb{X}) = 1$.*

Note that the ergodicity definition applies only to stationary sequences and that not all stationary sequences are ergodic. In fact, only those for which the set of path $\mathbb{X}$ is itself invariant to shifts qualify for the definition.

Example 1.13 *Consider a path on a unit circle. Let $\mathbb{X} = (y_0, \ldots, y_t)$ where each element of the sequence satisfies $y_j(\varkappa) = y_{j-1}(\ell\varkappa)$. Let $P(\mathbb{X})$ be the length of the interval $[y_0, y_t]$). Let $\ell^\tau\mathbb{X} = \{y_{o-\tau}, \ldots, y_{t-\tau}\}$ displace $\mathbb{X}$ by half a circle. Since $P(\ell^\tau\mathbb{X}) = P(\mathbb{X})$, y_t is stationary. However, $P(\ell^\tau\mathbb{X}) \neq 1$ or 0 so y_t is not ergodic.*

A weaker definition of ergodicity is the following:

Definition 1.13 *(Ergodicity 2): A (weakly) stationary process $\{y_t(\varkappa)\}$ is ergodic if and only if $\frac{1}{T}\sum_{t=1}^{T} y_t \overset{a.s.}{\to} E[y_t(\varkappa)]$, where the expectation is taken with respect to $\varkappa$.*

Definition 1.12 is stronger than definition 1.13 because it refers to the probability of paths (the latter concerns only their first moment). Intuitively, if a process is stationary its path converges to some limit. If it is stationary and ergodic, all paths (indexed by $\varkappa$) will converge to the same limit. Hence, one path is sufficient to infer the moments of its distribution.

Example 1.14 *Consider the process* $y_t = e_t - 2e_{t-1}$, *where* $e_t \sim$ *iid* $(0, \sigma_e^2)$. *It is easy to verify that* $E(y_t) = 0$, $var(y_t) = 5\sigma_e^2 < \infty$ *and* $cov(y_t, y_{t-\tau})$ *does not depend on* t. *Therefore the process is covariance stationary. To verify that it is ergodic consider the sample mean* $\frac{1}{T}\sum_t y_t$, *which is easily shown to converge to 0 as* $T \to \infty$. *The sample variance is* $\frac{1}{T}\sum_t y_t^2 = \frac{1}{T}\sum_t (e_t - 2e_{t-1})^2 = \frac{5}{T}\sum_t e_t^2$ *which converges to* $var(y_t)$ *as* $T \to \infty$.

Example 1.15 *Let* $y_t = e_1 + e_{2t}$ $t = 0, 1, 2, \ldots$, *where* $e_{2t} \sim$ *iid* $(0, 1)$ *and* $e_1 \sim (1, 1)$. *Clearly* y_t *is stationary and* $E(y_t) = 1$. *However,* $\frac{1}{T}\sum_t y_t = e_1 + \frac{1}{T}\sum_t e_{2t}$ *and* $\lim_{T\to\infty} \frac{1}{T}\sum_t y_t = e_1 + \lim_{T\to\infty} \frac{1}{T}\sum_t e_{2t} = e_1$, *because* $\frac{1}{T}\sum_t e_{2t} \overset{a.s.}{\to} 0$. *Since the time average of* y_t *(equal to* e_1*) is different from the population average of* y_t *(equal to* 1*),* y_t *is not ergodic.*

What is wrong with example 1.15? Intuitively, y_t is not ergodic because it has "too much" memory (e_1 appears in y_t for every t). In fact, for ergodicity to hold, the process must "forgets" its past reasonably fast.

Exercise 1.9 *Consider the process* $y_t = 0.6y_{t-1} + 0.2y_{t-2} + e_t$, *where* $e_t \sim$ *iid* $(0, 1)$. *Is* y_t *stationary? Is it ergodic? Find the effect of a unitary change in* e_t *on* y_{t+3}. *Repeat the exercise for* $y_t = 0.4y_{t-1} + 0.8y_{t-2} + e_t$.

Exercise 1.10 *Consider the bivariate process:*

$$\begin{aligned} y_{1t} &= 0.3y_{1t-1} + 0.8y_{2t-1} + e_{1t} \\ y_{2t} &= 0.3y_{1t-1} + 0.4y_{2t-1} + e_{2t} \end{aligned}$$

where $E(e_{1t}e_{1\tau}) = 1$ *for* $\tau = t$ *and 0 otherwise,* $E(e_{2t}e_{2\tau}) = 2$ *for* $\tau = t$ *and 0 otherwise, and* $E(e_{1t}e_{2\tau}) = 0$ *for all* τ, t. *Is the system covariance stationary? Is it ergodic? Calculate* $\frac{\partial y_{1t+\tau}}{\partial e_{2t}}$ *for* $\tau = 2, 3$. *What is the limit of this derivative as* $\tau \to \infty$*?*

Exercise 1.11 *Suppose that at* t *time 0,* $\{y_t\}_{t=1}^{\infty}$ *is given by*

$$y_t = \begin{cases} 1 & \textit{with probability } 1/2 \\ 0 & \textit{with probability } 1/2 \end{cases}$$

Show that y_t *is stationary but not ergodic. Show that a single path (i.e. a path composed of only 1's and 0's) is ergodic.*

Exercise 1.12 *Let $y_t = cos(\pi/2 \cdot t) + e_t$, where $e_t \sim$ iid $(0, \sigma_e^2)$. Show that y_t is neither stationary nor ergodic. Show that the sequence $\{y_t, y_{t+4}, y_{t+8} \ldots\}$ is stationary and ergodic.*

Exercise 1.12 shows an important result: if a process is non-ergodic, it may be possible to find a subsequence which is ergodic.

Exercise 1.13 *Show that if $\{y_{1t}(\varkappa)\}$ is ergodic, $y_{2t} = h(y_{1t})$ is ergodic if h is continuous.*

A concept which bears some resemblance with ergodicity is the one of mixing.

Definition 1.14 *(Mixing 1) Let $\mathbb{B}_1$ and $\mathbb{B}_2$ be two Borel algebra [3] and $B_1 \in \mathbb{B}_1$ and $B_2 \in \mathbb{B}_2$ two events. Then ϕ-mixing and α-mixing are defined as follows:*

$$\phi(\mathbb{B}_1, \mathbb{B}_2) \equiv \sup_{\{B_1 \in \mathbb{B}_1, B_2 \in \mathbb{B}_2 : P(B_1) > 0\}} |P(B_2|B_1) - P(B_2)|$$

$$\alpha(\mathbb{B}_1, \mathbb{B}_2) \equiv \sup_{\{B_1 \in \mathbb{B}_1, B_2 \in \mathbb{A}_2\}} |P(B_2 \cap B_1) - P(B_2)P(B_1)|.$$

Intuitively, ϕ-mixing and α-mixing measure the dependence of events. We say that events in $\mathbb{B}_1$ and $\mathbb{B}_2$ are independent if both ϕ and α are zero. The function ϕ provides a measure of relative dependence while α measures absolute dependence.

For a stochastic process α-mixing and ϕ-mixing are defined as follows. Let $\mathbb{B}_{-\infty}^t$ be the Borel algebra generated by values of y_t from the infinite past up to t and $\mathbb{B}_{t+\tau}^{\infty}$ be the Borel algebra generated by values of y_t from $t + \tau$ to infinity. Intuitively, $\mathbb{B}_{-\infty}^t$ contains information up to t and $\mathbb{B}_{t+\tau}^{\infty}$ information from $t + \tau$ on.

Definition 1.15 *(Mixing 2): For a stochastic process $\{y_t(\varkappa)\}$, the mixing coefficients ϕ and α are defined as: $\phi(\tau) = \sup_t \phi(\mathbb{B}_{-\infty}^t, \mathbb{B}_{t+\tau}^{\infty})$ and $\alpha(\tau) = \sup_t \alpha(\mathbb{B}_{-\infty}^j, \mathbb{B}_{t+\tau}^{\infty})$*

$\phi(\tau)$ and $\alpha(\tau)$, called respectively uniform and strong mixing, measure how much dependence there is between elements of $\{y_t\}$ separated by τ periods. If $\phi(\tau) = \alpha(\tau) = 0$, y_t and $y_{t+\tau}$ are independent. If $\phi(\tau) = \alpha(\tau) = 0$ as $\tau \to \infty$, they are asymptotically independent. Note that because $\phi(\tau) \geq \alpha(\tau)$, ϕ-mixing implies α-mixing.

Example 1.16 *Let y_t be such that $cov(y_t y_{t-\tau_1}) = 0$ for some τ_1. Then $\phi(\tau) = \alpha(\tau) = 0$, $\forall \tau \geq \tau_1$. Let $y_t = Ay_{t-1} + e_t$, $|A| \leq 1$, $e_t \sim$ iid $(0, \sigma_e^2)$. Then $\alpha(\tau) = 0$ as $\tau \to \infty$.*

Exercise 1.14 *Show that if $y_t = Ay_{t-1} + e_t$, $|A| \leq 1$, $e_t \sim$ iid $(0, \sigma_e^2)$, $\phi(\tau)$ does not go to zero as $\tau \to \infty$.*

[3] A Borel algebra is the smallest collection of subsets of the event space which allow us to express the probability of an event in terms of the sets of the algebra.

Mixing is a somewhat stronger memory requirement than ergodicity. Rosenblatt (1978) shows the following result:

Result 1.11 *Let y_t be stationary. If $\alpha(\tau) \to 0$ as $\tau \to \infty$, y_t is ergodic.* □

Exercise 1.15 *Use result 1.11 and the fact that $\phi(\tau) \geq \alpha(\tau)$ to show that if $\phi(\tau) \to 0$ as $\tau \to \infty$, a ϕ-mixing process is ergodic.*

Both ergodicity and mixing are hard to verify in practice. A concept which bears some relationship with both and it is easier to check is the following:

Definition 1.16 *(Asymptotic Uncorrelatedness): $y_t(\varkappa)$ has asymptotic uncorrelated elements if there exist constants $0 \leq \varrho_\tau \leq 1$, $\tau \geq 0$ such that $\sum_{\tau=0}^{\infty} \varrho_\tau < \infty$ and $cov(y_t, y_{t-\tau}) \leq \varrho_\tau \sqrt{(var(y_t) var(y_{t-\tau})}$, $\forall \tau > 0$, where $var(y_t) < \infty$, for all t.*

Intuitively, if we can find an upper bound to the correlation of y_t and $y_{t-\tau}$, $\forall \tau$, and if the accumulation over τ of this bound is finite, the process has a memory that asymptotically dies out.

Example 1.17 *Let $y_t = Ay_{t-1} + e_t$ $e_t \sim$ iid $(0, \sigma_e^2)$. Here $corr(y_t, y_{t-\tau}) = A^\tau$ and if $0 \leq A < 1$, $\sum_t A^\tau < \infty$, so that y_t has asymptotically uncorrelated elements.*

Note that in definition 1.16 only $\tau > 0$ matters. From example 1.17 it is clear that when $\text{var}(y_t)$ is constant and the covariance of y_t with $y_{t-\tau}$ only depends on τ, asymptotic uncorrelatedness is the same as covariance stationarity.

Exercise 1.16 *Show that for $\sum_{\tau=0}^{\infty} \varrho_\tau < \infty$ it is necessary that $\varrho_\tau \to 0$ as $\tau \to \infty$ and sufficient that $\rho_\tau < \tau^{-1-b}$ for some $b > 0$, τ sufficiently large.*

Exercise 1.17 *Suppose that y_t is such that the correlation between y_t and $y_{t-\tau}$ goes to zero as $\tau \to \infty$. Is this sufficient to ensure that y_t is ergodic?*

Instead of assuming stationarity and ergodicity or mixing, one can assume that y_t satisfies an alternative set of conditions. These conditions considerably broadens the set of time series a researcher can deal with.

Definition 1.17 *(Martingale): $\{y_t\}$ is a martingale with respect to the information set $\mathcal{F}_t$ if $y_t \in \mathcal{F}_t$ $\forall t > 0$ and $E_t[y_{t+\tau}] \equiv E[y_{t+\tau}|\mathcal{F}_t] = y_t$, for all t, τ.*

Definition 1.18 *(Martingale difference): $\{y_t\}$ is a martingale difference with respect to the information set $\mathcal{F}_t$ if $y_t \in \mathcal{F}_t$, $\forall t > 0$ and $E_t[y_{t+\tau}] \equiv E[y_{t+\tau}|F_t] = 0$ for all t, τ.*

Example 1.18 *Let y_t be iid with $E(y_t) = 0$. Let $\mathcal{F}_t = \{\ldots, y_{t-1}, y_t\}$ and let $\mathcal{F}_{t-1} \subseteq \mathcal{F}_t$. Then y_t is a martingale difference sequence.*

Martingale difference is a much weaker requirement than stationarity and ergodicity since it only involves restrictions on the first conditional moment. It is therefore easy to build examples of processes which are martingale difference but are not stationary.

Example 1.19 *Suppose that y_t is iid with mean zero and variance σ_t^2. Then y_t is a martingale difference, nonstationary process.*

Exercise 1.18 *Let y_{1t} be a stochastic process and let $y_{2t} = E[y_{1t}|\mathcal{F}_t]$ be its conditional expectation. Show that y_{2t} is a martingale.*

Using the identity $y_t = y_t - E(y_t|\mathcal{F}_{t-1}) + E(y_t|\mathcal{F}_{t-1}) - E(y_t|\mathcal{F}_{t-2}) + E(y_t|\mathcal{F}_{t-2})\ldots$, one can write $y_t = \sum_{j=0}^{\tau-1} Rev_{t-j}(t) + E(y_t|\mathcal{F}_{t-\tau})$ for $\tau = 1, 2, \ldots$ where $Rev_{t-j}(t) \equiv E[y_t|\mathcal{F}_{t-j}] - E[y_t|\mathcal{F}_{t-j-1}]$ is the one step ahead revision in y_t, made with new information accrued from $t-j-1$ to $t-j$. $Rev_{t-j}(t)$ plays an important role in deriving the properties of functions of stationary processes, and will be extensively used in chapters 4 and 10.

Exercise 1.19 *Show that $Rev_{t-j}(t)$ is a martingale difference.*

1.4 Law of Large Numbers

Laws of large numbers provide conditions to insure that quantities like $\frac{1}{T}\sum_t x_t'x_t$ or $\frac{1}{T}\sum_t z_t'x_t$, which appear in the formulas of linear estimators like OLS or IV, stochastically converge to well defined limits. Since different conditions apply to different kinds of economic data, we consider here situations which are typically encountered in macro-time series contexts. Given the results of section 2, we will describe only strong law of large numbers since weak law of large numbers hold as a consequence.

Laws of large numbers typically come in the following form: given restrictions on the dependence and the heterogeneity of the observations and/or some moments restrictions, $\frac{1}{T}\sum y_t - E(y_t) \overset{a.s.}{\to} 0$. We will consider three cases: (i) y_t has dependent and identically distributed elements, (ii) y_t has dependent and heterogeneously distributed elements, (iii) y_t has martingale difference elements. To better understand the applicability of each case note that in all cases observations are serially correlated. In the first case we restrict the distribution of the observations to be the same for every t; in the second we allow some carefully selected form of heterogeneity (for example, structural breaks in the mean or in the variance or conditional heteroschedasticity); in the third we do not restrict the distribution of the process, but impose conditions on its moments.

1.4.1 Dependent and Identically Distributed Observations

To state a law of large numbers (LLN) for stationary processes we need conditions on the memory of the sequence. Typically, one assumes ergodicity since this implies average asymptotic independence of the elements of the $\{y_t(\varkappa)\}$ sequence.

The LLN is then as follows: Let $\{y_t(\varkappa)\}$ be stationary and ergodic with $E|y_t| < \infty \ \forall t$. Then $\frac{1}{T}\sum_t y_t \overset{a.s.}{\to} E(y_t)$ (See Stout, 1974, p. 181). □

To use this law when dealing with econometric estimators recall that for any measurable function h such that $y_{2t} = h(y_{1t})$, y_{2t} is stationary and ergodic if y_{1t} is stationary and ergodic.

Exercise 1.20 *(Strong consistency of OLS and IV estimators): Let $y_t = x_t\alpha_0 + e_t$; let $x = [x_1 \cdots x_T]'$, $z = [z_1, \cdots z_T]'$, $e = [e_1, \cdots, e_T]'$ and assume:*
(i) $\frac{x'e}{T} \overset{a.s.}{\to} 0$
(i') $\frac{z'e}{T} \overset{a.s.}{\to} 0$
(ii) $\frac{x'x}{T} \overset{a.s.}{\to} \Sigma_{xx}$, Σ_{xx} finite, $|\Sigma_{xx}| \neq 0$
(ii') $\frac{z'x}{T} \overset{a.s.}{\to} \Sigma_{zx}$, Σ_{zx} finite, $|\Sigma_{zx}| \neq 0$
(ii") $\frac{z'x}{T} - \Sigma_{zx,T} \overset{a.s.}{\to} 0$, where $\Sigma_{zx,T}$ is $O(1)$ random matrix which depends on T and has uniformly continuous column rank.
Show that $\alpha_{OLS} = (x'x)^{-1}(x'y)$ and $\alpha_{IV} = (z'x)^{-1}(z'y)$ exist almost surely for T large and that $\alpha_{OLS} \overset{a.s.}{\to} \alpha_0$ under (i)-(ii) and that $\alpha_{IV} \overset{a.s.}{\to} \alpha_0$ under (i')-(ii'). Show that under (i)-(ii") α_{IV} exists almost surely for T large, and $\alpha_{IV} \overset{a.s.}{\to} \alpha_0$. (Hint: If A_n is a sequence of $k_1 \times k$ matrices, then A_n has uniformly full column rank if there exist a sequence of $k \times k$ submatrices Δ_n which is uniformly nonsingular.)

1.4.2 Dependent and Heterogeneously Distributed Observations

To derive a LLN for dependent and heterogeneously distributed processes we drop the ergodicity assumption and we substitute it with a mixing requirement. In addition, we need to define the *size* of the mixing conditions:

Definition 1.19 *Let $1 \leq a \leq \infty$. Then $\phi(\tau) = O(\tau^{-b})$ for $b > a/(2a-1)$ implies that $\phi(\tau)$ is of size $a/(2a-1)$. If $a > 1$ and $\alpha(\tau) = O(\tau^{-b})$ for $b > a/(a-1)$, $\alpha(\tau)$ is of size $a/(a-1)$.*

With definition 1.19 one can make precise statements on the memory of the process. Roughly speaking, the memory of a process is related to a. As $a \to \infty$, the dependence increases while as $a \to 1$, the sequence exhibits less and less serial dependence.

The LLN is the following. Let $\{y_t(\varkappa)\}$ be a sequence with $\phi(\tau)$ of size $a/(2a-1)$ or $\alpha(\tau)$ of size $a/(a-1)$, $a > 1$ and $E(y_t) < \infty$, $\forall t$. If for some $0 < b \leq a$, $\sum_{t=1}^{\infty}(\frac{E|y_t - E(y_t)|^{a+b}}{t^{a+b}})^{\frac{1}{a}} < \infty$, then $\frac{1}{T}\sum_t y_t - E(y_t) \overset{a.s.}{\to} 0$. (see McLeish, 1975, theorem 2.10). □.

Note that in the above law, the elements of y_t are allowed to have distributions that vary over time (e.g. $E(y_t)$ may depend on t) but the condition $(\frac{E|y_t - (E(y_t)|^{a+b}}{t^{a+b}})^{\frac{1}{a}} < \infty$ restricts the moments of the process. Note that for $a = 1$ and $b = 1$, the above collapses to Kolmogorov law of large numbers.
The moment condition can be weakened somewhat if we are willing to impose a bound on the $(a+b)$-th moment.

Result 1.12 *Let $\{y_t(\varkappa)\}$ be a sequence with $\phi(\tau)$ of size $a/(2a-1)$ or $\alpha(\tau)$ of size $a/(a-1)$, $a>1$ such that $E|y_t|^{a+b}$ is bounded for all t. Then $\frac{1}{T}\sum_t y_t - E(y_t) \xrightarrow{a.s.} 0$.* □

The next result mirrors the one obtained for stationary ergodic processes.

Result 1.13 *Let h be a measurable function and $y_{2\tau} = h(y_{1t}, \dots y_{1t+\tau})$, τ finite. If y_{1t} is mixing such that $\phi(\tau)$ $(\alpha(\tau))$ is $O(\tau^{-b})$ for some $b>0$, $y_{2\tau}$ is mixing such that $\phi(\tau)$ $(\alpha(\tau))$ is $O(\tau^{-b})$.* □

From the above result it immediately follows that if $\{z_t, x_t, e_t\}$ is a vector of mixing processes, $\{x_t'x_t\}, \{x_t'e_t\}$, $\{z_t'x_t\}, \{z_t'e_t\}$, are also mixing processes of the same size.

A useful result when observations are heterogeneous is the following:

Result 1.14 *Let $\{y_t(\varkappa)\}$ be a such that $\sum_{t=1}^{\infty} E|y_t| < \infty$. Then $\sum_{t=1}^{\infty} y_t$ converges almost surely and $E(\sum_{t=1}^{\infty} y_t) = \sum_{t=1}^{\infty} E(y_t) < \infty$ (see White, 1984, p.48).* □

A LLN for processes with asymptotically uncorrelated elements is the following. Let $\{y_t(\varkappa)\}$ be a process with asymptotically uncorrelated elements, mean $E(y_t)$, variance $\sigma_t^2 < \Delta < \infty$. Then $\frac{1}{T}\sum_t y_t - E(y_t) \xrightarrow{a.s.} 0$.

Compared with result 1.12, we have relaxed the dependence restriction from mixing to asymptotic uncorrelation at the cost of altering the restriction on moments of order $a+b$ $(a \geq 1, b \leq a)$ to second moments. Note that since functions of asymptotically uncorrelated processes are not asymptotically uncorrelated, to prove consistency of econometric estimators when the regressors have asymptotic uncorrelated increments we need to make assumptions on quantities like $\{x_t'x_t\}, \{x_t'e_t\}$, etc. directly.

1.4.3 Martingale Difference Process

A LLN for this type of processes is the following: Let $\{y_t(\varkappa)\}$ be a martingale difference. If for some $a \geq 1, \sum_{t=1}^{\infty} \frac{E|y_t|^{2a}}{t^{1+a}} < \infty$, then $\frac{1}{T}\sum_t y_t \xrightarrow{a.s.} 0$.

The martingale LLN requires restrictions on the moments of the process which are slightly stronger than those assumed in the case of independent y_t. The analogous of result 1.12 for martingale differences is the following.

Result 1.15 *Let $\{y_t(\varkappa)\}$ be a martingale difference such that $E|y_t|^{2a} < \Delta < \infty$, some $a \geq 1$ and all t. Then $\frac{1}{T}\sum_t y_t \xrightarrow{a.s.} 0$.* □

Exercise 1.21 *Suppose $\{y_{1t}(\varkappa)\}$ is a martingale difference. Show that $y_{2t} = y_{1t}z_t$ is a martingale difference for any $z_t \in \mathcal{F}_t$.*

Exercise 1.22 *Let $y_t = x_t\alpha_0 + e_t$, and assume (i) e_t is a martingale difference; (ii) $E(x_t'x_t)$ is positive and finite. Show that α_{OLS} exists and $\alpha_{OLS} \xrightarrow{a.s.} \alpha_0$.*

1.5 Central Limit Theorems

There are also several central limit theorems (CLT) available in the literature. Clearly, their applicability depends on the type of data a researcher has available. In this section we list CLTs for the three cases we have described in section 4. Loeve (1977) or White (1984) provide theorems for other relevant cases.

1.5.1 Dependent and Identically Distributed Observations

A central limit theorem for dependent and identically distributed observations can be obtained using two conditions. First, we need a restriction on the variance of the process. Second, we need to impose $E(y_t|\mathcal{F}_{t-\tau}) \to 0$ for $\tau \to \infty$ (referred as linear regularity in chapter 4) or $E[y_t|\mathcal{F}_{t-\tau}] \xrightarrow{q.m.} 0$ as $\tau \to \infty$. The second condition is obviously stronger than the first one. Restrictions on the variance of the process are needed since when y_t is a dependent and identically distributed process its variance is the sum of the variances of the forecast revisions made at each t, and this may not converges to a finite limit. We ask the reader to show this in the next two exercises.

Exercise 1.23 *Let* $var(y_t) = \sigma_y^2 < \infty$. *Show that* $cov(Rev_{t-j}(t), Rev_{t-j'}(t)) = 0,\ j < j'$, *where* $Rev_{t-j}(t)$ *was defined right before exercise 1.19. Note that this implies that* $\sigma_y^2 = var(\sum_{j=0}^{\infty} Rev_{t-j}(t)) = \sum_{j=0}^{\infty} var(Rev_{t-j}(t))$.

Exercise 1.24 *Let* $\overline{\sigma}_T^2 = T \times E((T^{-1}\sum_{t=1}^{T} y_t)^2)$. *Show that* $\overline{\sigma}_T^2 = \sigma_y^2 + 2\sigma_y^2 \sum_{\tau=1}^{T-1} \rho_\tau(1 - \tau/T)$, *where* $\rho_\tau = E(y_t y_{t-\tau})/\sigma_y^2$. *Give conditions on* y_t *that make* ρ_τ *independent of* t. *Show that* $\overline{\sigma}_T^2$ *grows without bound as* $T \to \infty$.

A sufficient condition insuring that $\overline{\sigma}_T^2$ converges is that $\sum_{j=0}^{\infty}(varRev_{t-j}(t))^{1/2} < \infty$.

A CLT is then as follows: Let (i) $\{y_t(\varkappa)\}$ be stationary and ergodic process, $y_t \in \mathcal{F}_t\ \forall t > 0$; (ii) $E(y_t^2) = \sigma_y^2 < \infty$; (iii) $E(y_t|F_{t-\tau}) \xrightarrow{q.m.} 0$ as $\tau \to \infty$; (iv) $\sum_{j=0}^{\infty}(varRev_{t-j}(t))^{1/2} < \infty$. Then, as $T \to \infty,\ 0 \neq \overline{\sigma}_T^2 \longrightarrow \bar{\sigma}_y^2 < \infty$ and $\sqrt{T}\frac{(\frac{1}{T}\sum_t y_t)}{\overline{\sigma}_T} \xrightarrow{D} \mathbb{N}(0,1)$ (see Gordin, 1969). □

Example 1.20 *An interesting pathological case obtains when* $\bar{\sigma}_T^2 = 0$. *Consider for example* $y_t = e_t - e_{t-1}, e_t \sim iid(0, \sigma_e^2)$. *Then* $\bar{\sigma}_T^2 = 2\sigma_e^2 - 2\sigma_e^2 = 0$. *Hence* $\frac{1}{T}\sum_t y_t = \frac{1}{T}(y_t - y_0)$ *and* $\sqrt{T}(\frac{1}{T}\sum_t y_t) \xrightarrow{P} 0$.

Exercise 1.25 *Assume that (i)* $E[x_{tji}e_{tj}|\mathcal{F}_{t-1}] = 0\ \forall t\ \ i = 1, \ldots;\ j = 1, \ldots$; *(ii)* $E[x_{tji}e_{tj}]^2 < \infty$; *(iii)* $\Sigma_T \equiv var(T^{-1/2}x'e) \to var(x'e) \equiv \Sigma$ *as* $T \to \infty$ *is nonsingular and positive definite; (iv)* $\sum_j (varRev_{t-j}(t))^{-1/2} < \infty$; *(v)* (x_t, e_t) *are stationary ergodic sequences; (vi)* $E|x_{tji}|^2 < \infty$; *(vii)* $\Sigma_{xx} \equiv E(x_t' x_t)$ *is positive definite. Show that* $(\Sigma_{xx}^{-1}\Sigma\Sigma_{xx}^{-1'})^{-1/2}\sqrt{T}(\alpha_{OLS} - \alpha_0) \overset{D}{\sim} \mathbb{N}(0, I)$ *where* α_{OLS} *is the OLS estimator of* α_0 *in the model* $y_t = x_t\alpha_0 + e_t$ *and* T *is the number of observations.*

1.5.2 Dependent Heterogeneously Distributed Observations

The CLT in this case is the following: Let $\{y_t(\varkappa)\}$ be a sequence of mixing random variables such that either $\phi(\tau)$ or $\alpha(\tau)$ is of size $a/a-1,\ a>1$ with $E(y_t)=0$; and $E|y_t|^{2a}<\Delta<\infty,\ \forall t$. Define $y_{b,T}=\frac{1}{\sqrt{T}}\sum_{t=b+1}^{b+T}y_t$ and assume there exists a $0\neq\overline{\sigma}^2<\infty$, such that $E(y_{b,T}^2)\longrightarrow\overline{\sigma}^2$ for $T\to\infty$, uniformly in b. Then $\sqrt{T}\frac{(\frac{1}{T}\sum_t y_t)}{\overline{\sigma}_T}\xrightarrow{D}\mathbb{N}(0,1)$, where $\overline{\sigma}_T^2\equiv E(y_{0,T}^2)$ (see White and Domowitz (1984)). □

As in the previous CLT, we need the condition that the variance of y_t is consistently estimated. Note also that we have substituted stationarity and ergodicity assumptions with the one of mixing and that we need uniform convergence of $E(y_{b,T}^2)$ to $\overline{\sigma}^2$ in b. This is equivalent to imposing that y_t is asymptotically covariance stationary, that is, that heterogeneity in y_t dies out at T increases (see White, 1984, p.128).

1.5.3 Martingale Difference Observations

The CLT in this case is as follows: Let $\{y_t(\varkappa)\}$ be a martingale difference process with $\sigma_t^2\equiv E(y_t^2)<\infty,\ \sigma_t^2\neq 0,\ \mathcal{F}_{t-1}\subset\mathcal{F}_t,\ y_t\in\mathcal{F}_t$; let $\mathcal{D}_t$ be the distribution function of y_t and let $\overline{\sigma}_T^2=\frac{1}{T}\sum_{t=1}^T\sigma_t^2$. If for every $\epsilon>0$, $\lim_{T\to\infty}\overline{\sigma}_T^{-2}\frac{1}{T}\sum_{t=1}^T\int_{y^2>\epsilon T\overline{\sigma}_T^2}y^2d\mathcal{D}_t(y)=0$ and $(\frac{1}{T}\sum_{t=1}^T y_t^2)/\overline{\sigma}_T^2-1\xrightarrow{P}0$ then $\sqrt{T}\frac{(\frac{1}{T}\sum_t y_t)}{\overline{\sigma}_T}\xrightarrow{D}\mathbb{N}(0,1)$ (See McLeish, 1974). □

The last condition is somewhat mysterious: it requires that the average contribution of the extreme tails of the distribution to the variance of y_t is zero in the limit. If this condition holds then y_t satisfies a uniform asymptotic negligibility condition. In other words, none of the elements of $\{y_t(\varkappa)\}$ can have a variance which dominates the variance of $\frac{1}{T}\sum_t y_t$. We illustrate this condition in the next example.

Example 1.21 *Suppose* $\sigma_t^2=\rho^t,\ 0<\rho<1$. *Then* $T\overline{\sigma}_T^2\equiv\sum_{t=1}^T\sigma_t^2=\sum_{t=1}^T\rho^t=\frac{\rho}{1-\rho}$ *as* $T\to\infty$. *In this case* $\max_{1\leq t\leq T}\frac{\sigma_t^2}{T\overline{\sigma}_T^2}=\rho/\frac{\rho}{1-\rho}=1-\rho\neq 0$, *independent of* T. *Hence the asymptotic negligibility condition is violated. Let now* $\sigma_t^2=\sigma^2$, $\overline{\sigma}_T^2=\sigma^2$. *Then* $\max_{1\leq t\leq T}\frac{\sigma_t^2}{T\overline{\sigma}_T^2}=\frac{1}{T}\frac{\sigma^2}{\sigma^2}\longrightarrow 0$ *as* $T\to\infty$ *and the asymptotic negligibility condition holds.*

The martingale difference assumption allows us to weaken several of the conditions needed to prove a central limit theorem relative to the case of stationary processes and will be the one used in several parts of this book.

A result, which we will become useful in later chapters concerns the asymptotic distribution of functions of converging stochastic processes.

Result 1.16 *Suppose the* $m\times 1$ *vector* $\{y_t(\varkappa)\}$ *is asymptotically normally distributed with mean* $\bar{y}$ *and variance* $a_t^2\Sigma_y$, *where* Σ_y *is a symmetric, non-negative definite matrix and* $a_t\to 0$ *as* $t\to\infty$. *Let* $h(y)=(h_1(y),\ldots,h_n(y))'$ *be such that each* $h_j(y)$ *is continuously differentiable in the neighborhood of* $\bar{y}$ *and let* $\Sigma_h=\frac{\partial h(\bar{y})}{\partial y'}\Sigma_y(\frac{\partial h(\bar{y})}{\partial y'})'$ *have nonzero diagonal elements, where* $\frac{\partial h(\bar{y})}{\partial y'}$ *is a* $n\times m$ *matrix. Then* $h(y_t)\xrightarrow{D}\mathbb{N}(h(\bar{y}),a_t^2\Sigma_h)$. □

Example 1.22 *Suppose y_t is iid with mean $\bar{y}$ and variance σ_y^2, $\bar{y} \neq 0, 0 < \sigma_y^2 < \infty$. Then by the CLT $\frac{1}{T}\sum_t y_t \xrightarrow{D} N(\bar{y}, \frac{\sigma_y^2}{T})$ and by result 1.16 $(\frac{1}{T}\sum_t y_t)^{-1} \xrightarrow{D} N(\bar{y}^{-1}, \frac{\sigma_y^2}{T\bar{y}^4})$.*

1.6 Elements of Spectral Analysis

A central object in the analysis of time series is the spectral density.

Definition 1.20 *(Spectral density): The spectral density of stationary $\{y_t(\varkappa)\}$ process at frequency $\omega \in [0, 2\pi]$ is $\mathcal{S}_y(\omega) = \frac{1}{2\pi}\sum_{\tau=-\infty}^{\infty} ACF_y(\tau)\exp\{-i\omega\tau\}$.*

We have already mentioned that the spectral density is a reparametrization of the covariance generating function and it is obtained setting $z = e^{-i\omega} = cos(\omega) - isin(\omega)$, where $i = \sqrt{-1}$. Definition 1.20 also shows that the spectral density is the Fourier transform of the autocovariance of y_t. Hence, the spectral density simply repackages the autocovariances of $\{y_t(\varkappa)\}$ using sine and cosine functions as weights but at times it is more useful than the autocovariance function since, for ω appropriately chosen, its elements are uncorrelated.

Example 1.23 *Two elements of the spectral density typically of interest are $\mathcal{S}(\omega) = 0$ and $\sum_j \mathcal{S}(\omega_j)$. It is easily verified that $\mathcal{S}(\omega = 0) = \frac{1}{2\pi}\sum_\tau ACF(\tau) = \frac{1}{2\pi}(ACF(0) + 2\sum_{\tau=1}^{\infty} ACF(\tau))$, that is, the spectral density at frequency zero is the (unweighted) sum of all the elements of the autocovariance function. It is also easy to verify that $\sum_j \mathcal{S}(\omega_j) = var(y_t)$, that is, the variance of the process is the area below the spectral density.*

To understand how the spectral density transforms the autocovariance function select, for example, $\omega = \frac{\pi}{2}$. Note that $\cos(\frac{\pi}{2}) = 1$, $cos(\frac{3\pi}{2}) = -1$, $\cos(\pi) = \cos(2\pi) = 0$ and that $\sin(\frac{\pi}{2}) = sin(\frac{3\pi}{2}) = 0$, $\sin(0) = 1$ and $\sin(\pi) = -1$ and that these values repeat themselves since the sine and cosine functions are periodic.

Exercise 1.26 *Calculate $S(\omega = \pi)$. Which autocovariances enter at frequency π?*

It is typical to evaluate the spectral density at Fourier frequencies i.e. at $\omega_j = \frac{2\pi j}{T}$, $j = 1, \ldots, T-1$, since for any two $\omega_1 \neq \omega_2$ such frequencies, $\mathcal{S}(\omega_1)$ is uncorrelated with $\mathcal{S}(\omega_2)$. Note that Fourier frequencies change with T, making recursive evaluation of the spectral density cumbersome. For a Fourier frequency, the period of oscillation is $\frac{2\pi}{\omega_j} = \frac{T}{j}$.

Example 1.24 *Suppose you have quarterly data. Then at the Fourier frequency $\frac{\pi}{2}$, the period is equal to 4. That is, at frequency $\frac{\pi}{2}$ you have fluctuations with an annual periodicity. Similarly, at the frequency π, the period is 2 so that biannual cycles are present at π.*

Exercise 1.27 *Business cycle are typically thought to occur with a periodicity between 2 and 8 years. Assuming that you have quarterly data, find the Fourier frequencies characterizing business cycle fluctuations. Repeat the exercise for annual and monthly data.*

Given the formula to calculate the period of oscillation, it is immediate to note that low frequencies are associated with cycles of long periods of oscillation - that is, with infrequent shifts from a peak to a through - and high frequencies with cycles of short periods of oscillation - that is, with frequent shifts from a peak to a through (see figure 1.1). Hence, trends (i.e. cycles with an infinite periodicity) are located in the low frequencies of the spectrum and irregular fluctuations in the high frequencies. Since the spectral density is periodic $mod(2\pi)$ and symmetric around $\omega = 0$, it is sufficient to examine $\mathcal{S}(\omega)$ over the interval $[0, \pi]$.

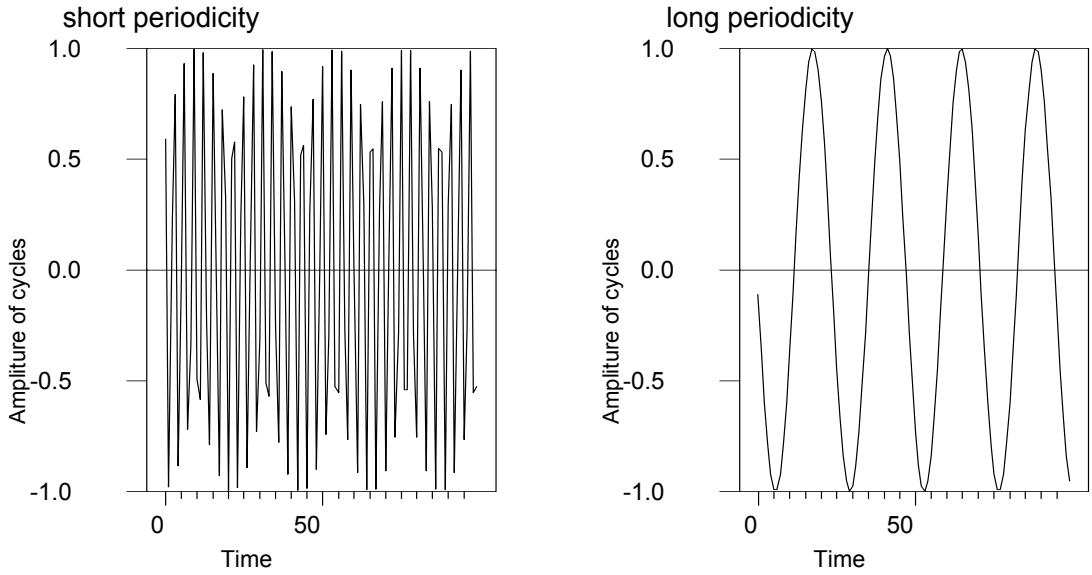


Figure 1.1: Short and long cycles

Exercise 1.28 *Show that* $\mathcal{S}(\omega_j) = \mathcal{S}(-\omega_j)$.

Example 1.25 *Suppose* $\{y_t(\varkappa)\}$ *is iid* $(0, \sigma_y^2)$. *Then* $ACF_y(\tau) = \sigma_y^2$ *for* $\tau = 0$ *and zero otherwise and* $\mathcal{S}_y(\omega_j) = \frac{\sigma^2}{2\pi}$, $\forall \omega_j$. *That is, the spectral density of an iid process is constant for all* $\omega_j \in [0, \pi]$.

Exercise 1.29 *Consider a stationary AR(1) process* $\{y_t(\varkappa)\}$ *with autoregressive coefficient equal to* $0 \leq A < 1$. *Calculate the autocovariance function of* y_t. *Show that the spectral density is monotonically increasing as* $\omega_j \to 0$.

Exercise 1.30 *Consider a stationary MA(1) process* $\{y_t(\varkappa)\}$ *with MA coefficient equal to* D*. Calculate the autocovariance function and the spectral density of* y_t*. Show its shape when* $D > 0$ *and when* $D < 0$*.*

Economic time series have a typical bell shaped spectral density (see figure 1.2) with a large portion of the variance concentrated in the lower part of the spectrum. Given the result of exercise 1.29, it is therefore reasonable to posit that most of economic time series can be represented with relatively simple AR processes.

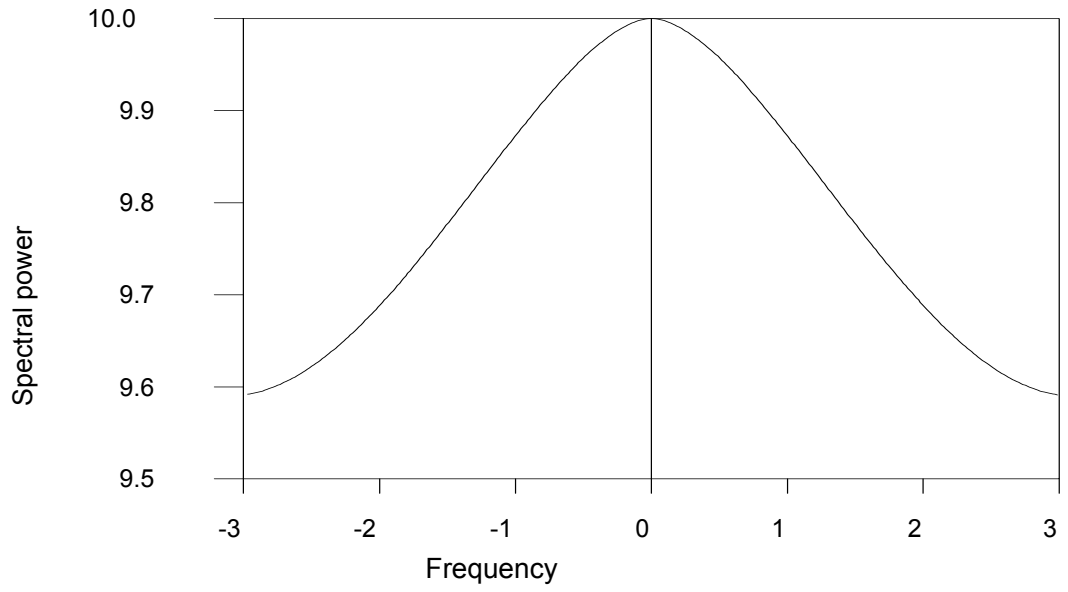


Figure 1.2: Spectral Density

The definitions we have given are valid for univariate processes but can be easily extended to vector of stochastic processes.

Definition 1.21 *(Spectral density matrix): The spectral density matrix of an* $m \times 1$ *vector of stationary processes* $\{y_t(\varkappa)\}$ *is* $\mathcal{S}_y(\omega) = \frac{1}{2\pi} \sum_\tau ACF_y(\tau) e^{-i\omega\tau}$ *where*

$$\mathcal{S}_y(\omega) = \begin{bmatrix} \mathcal{S}_{y_1y_1}(\omega) & \mathcal{S}_{y_1y_2}(\omega) & \dots & \mathcal{S}_{y_1y_m}(\omega) \\ \mathcal{S}_{y_2y_1}(\omega) & \mathcal{S}_{y_2y_2}(\omega) & \dots & \mathcal{S}_{y_2y_m}(\omega) \\ \dots & \dots & \dots & \dots \\ \mathcal{S}_{y_my_1}(\omega) & \mathcal{S}_{y_my_2}(\omega) & \dots & \mathcal{S}_{y_my_m}(\omega) \end{bmatrix}$$

The elements on the diagonal of the spectral density matrix are real while the elements off-the diagonal are typically complex. A measure of the strength of the relationship between two series at frequency ω is given by the coherence.

Definition 1.22 *Consider a bivariate stationary process* $\{y_{1t}(\varkappa), y_{2t}(\varkappa)\}$. *The coherence between* $\{y_{1t}(\varkappa)\}$ *and* $\{y_{2t}(\varkappa)\}$ *at frequency* ω_j *is* $Co(\omega) = \frac{|\mathcal{S}_{y_1,y_2}(\omega)|}{\sqrt{\mathcal{S}_{y_1,y_1}(\omega)\mathcal{S}_{y_2,y_2}(\omega)}}$.

The coherence is the frequency domain version of the correlation coefficient. Notice that $Co(\omega)$ is a real valued function where $|y|$ indicates the real part (or the modulus) of complex number y.

Example 1.26 *Suppose* $y_t = D(\ell)e_t$, *where* $e_t \sim$ *iid* $(0, \sigma_e^2)$. *It is immediate to verify that the coherence between* e_t *and* y_t *is one at all frequencies. Suppose, on the other hand, that* $Co(\omega)$ *monotonically declines to zero as* ω *moves from 0 to* π. *Then* y_t *and* e_t *have similar low frequency but different high frequency components.*

Exercise 1.31 *Suppose that* $e_t \sim$ *iid* $(0, \sigma_e^2)$ *and let* $y_t = Ay_{t-1} + e_t$. *Calculate* $Co_{y_t,e_t}(\omega)$.

Interesting transformations of y_t can be obtained with the use of filters.

Definition 1.23 *A filter is a linear transformation of a stochastic process, i.e. if* $y_t = \mathcal{B}(\ell)e_t$, $e_t \sim$ *iid* $(0, \sigma_e^2)$, *then* $\mathcal{B}(\ell)$ *is a filter.*

A moving average (MA) process is therefore a filter since a white noise is linearly transformed into another process. In general, stochastic processes can be thought of as filtered versions of some white noise process (the news). To study the spectral properties of filtered processes let $CGF_e(z)$ be the covariance generating function of e_t. Then the covariance generating function of y_t is $CGF_y(z) = \mathcal{B}(z)\mathcal{B}(z^{-1})CGF_e(z) = |\mathcal{B}(z)|^2 CGF_e(z)$, where $|\mathcal{B}(z)|$ is the modulus of $\mathcal{B}(z)$.

Example 1.27 *Suppose that* $e_t \sim$ *iid* $(0, \sigma_e^2)$ *so that its spectrum is* $\mathcal{S}_e(\omega) = \frac{\sigma^2}{2\pi}, \forall\omega$. *Consider now the process* $y_t = D(\ell)e_t$, *where* $D(\ell) = D_0 + D_1\ell + D_2\ell^2 + \ldots$. *It is typical to interpret* $D(\ell)$ *as the response function of* y_t *to a unitary change in* e_t. *Then* $\mathcal{S}_y(\omega) = |D(e^{-i\omega})|^2\mathcal{S}_e(\omega)$, *where* $|D(e^{-i\omega})|^2 = D(e^{-i\omega})D(e^{i\omega})$ *and* $D(e^{-i\omega}) = \sum_\tau D_\tau e^{-i\omega\tau}$ *measures how a unitary change in* e_t *affects* y_t *at frequency* ω.

Example 1.28 *Suppose that* $y_t = \bar{y} + at + D(\ell)e_t$, *where* $e_t \sim$ *iid* $(0, \sigma_e^2)$. *Since* y_t *displays a (linear) trend is not stationary and* $\mathcal{S}(\omega)$ *does not exist. Differencing the process we have* $y_t - y_{t-1} = a + D(\ell)(e_t - e_{t-1})$ *so that* $y_t - y_{t-1}$ *is stationary if* $e_t - e_{t-1}$ *is a stationary and all the roots of* $D(\ell)$ *are greater than one in absolute value. If these conditions are met, the spectrum of* Δy_t *is well defined and equals* $\mathcal{S}_{\Delta y}(\omega) = |D(e^{-i\omega})|^2\mathcal{S}_{\Delta e}(\omega)$.

The quantity $\mathcal{B}(e^{-i\omega})$ is called transfer function of the filter. Various functions of this quantity are of interest. For example, $|\mathcal{B}(e^{-i\omega})|^2$, the square modulus of the transfer function, measures the change in variance of e_t induced by the filter. Furthermore, since $\mathcal{B}(e^{-i\omega})$ is complex two alternative representations of the transfer function exist. The first decomposes it into its real and complex part, i.e. $\mathcal{B}(e^{-i\omega}) = \mathcal{B}^{\dagger}(\omega) + i\mathcal{B}^{\ddagger}(\omega)$, where both $\mathcal{B}^{\dagger}$ and $\mathcal{B}^{\ddagger}$ are real. Then the phase shift $Ph(\omega) = \tan^{-1}[\frac{-\mathcal{B}^{\ddagger}(\omega)}{\mathcal{B}^{\dagger}(\omega)}]$, measures how much the lead-lag relationships in e_t are altered by the filter. The second can be obtained using the polar representation $\mathcal{B}(e^{-i\omega}) = Ga(\omega)e^{-iPh(\omega)}$, where $Ga(\omega)$ is the gain. Here $Ga(\omega) = |\mathcal{B}(e^{-i\omega})|$ measures the change in the amplitude of cycles induced by the filter.

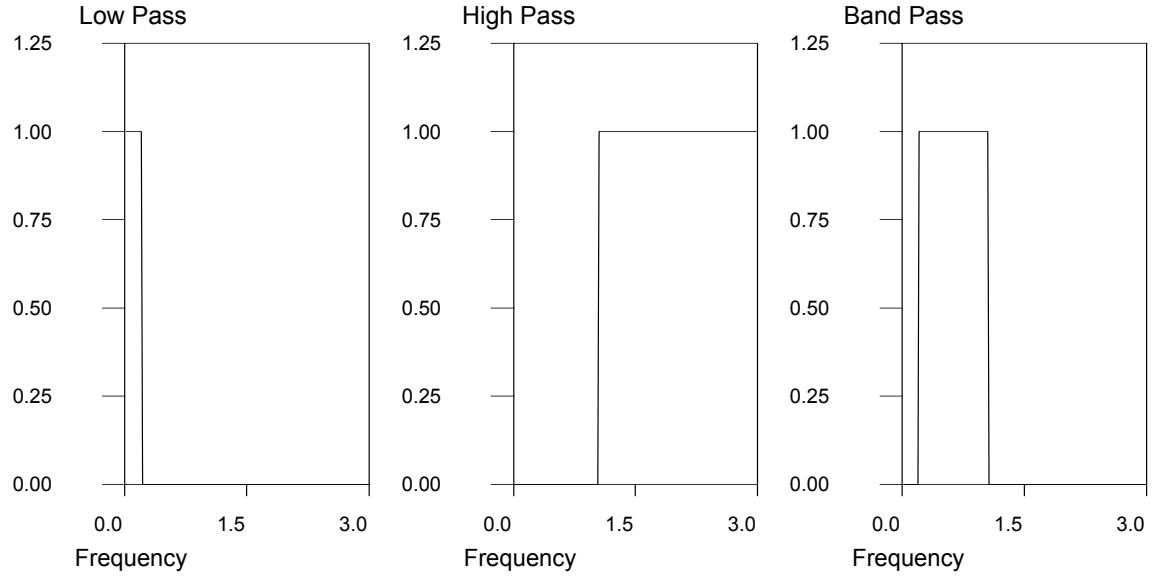


Figure 1.3: Filters

Filtering is an operation frequently performed in every day life (e.g. tuning a radio on a station filters out all other signals (waves)). Several types of filters are used in modern macroeconomics. Figure 1.3 presents three general types of filters: a low pass, a high pass, and a band pass. A low pass filter leaves the low frequencies of the spectrum unchanged but wipes out high frequencies. A high pass filter does exactly the opposite. A band pass filter can be thought as a combination of a low pass and a high pass filters: it wipes out very high and very low frequencies and leaves unchanged frequencies in middle range.

Low pass, high pass and band pass filters are non-realizable, in the sense that with samples of finite length, it is impossible to construct objects that looks like those of figure 1.3. In fact, using the inverse Fourier transform, one can show that these three filters (denoted, respectively, by $\mathcal{B}(\ell)^{lp}, \mathcal{B}(\ell)^{hp}, \mathcal{B}(\ell)^{bp}$) have the time representation:

- Low pass: $\mathcal{B}_0^{lp} = \frac{\omega_1}{\pi}$; $\mathcal{B}_j^{lp} = \frac{sin(j\omega_1)}{j\pi}$; $\forall j > 0$, some $\omega_1 \in (0, \pi)$.
- High pass: $\mathcal{B}_0^{hp} = 1 - \mathcal{B}_0^{lp}$; $\mathcal{B}_j^{hp} = -\mathcal{B}_j^{lp}$; $\forall j > 0$.

• Band pass: $\mathcal{B}_j^{bp} = \mathcal{B}_j^{lp}(\omega_2) - \mathcal{B}_j^{lp}(\omega_1); \quad \forall j > 0,\ \omega_2 > \omega_1.$

When j is finite the box-like spectral shape of these filters can only be approximated with a bell-shaped function. This means that relative to the ideal, realizable filters generate a loss of power at the edges of the band (a phenomenon called leakage) and an increase in the importance of the frequencies in the middle of the band (a phenomenon called compression). Approximations to these ideal filters are discussed in chapter 3.

Definition 1.24 *The periodogram of a stationary $y_t(\varkappa)$ is $Pe_y(\omega) = \sum_\tau \widehat{ACF}(\tau)e^{-i\omega\tau}$, where $\widehat{ACF}_y = \frac{1}{T}\sum_t (y_t - \frac{1}{T}\sum_t y_t)(y_{t-\tau} - \frac{1}{T}\sum_t y_{t-\tau})'$.*

Perhaps surprisingly, the periodogram is an inconsistent estimator of the spectrum (see e.g. Priestley (1981, p. 433)). Intuitively, this occurs because it consistently captures the power of y_t over a band of frequencies but not in each single one of them. To obtain consistent estimates it is necessary to "smooth" periodogram estimates with a filter. Such a smoothing filter is typically called a "kernel".

Definition 1.25 *For any $\epsilon > 0$, a filter $\mathcal{B}(\omega)$ is a kernel (denoted by $\mathcal{K}_T(\omega)$) if $\mathcal{K}_T(\omega) \to 0$ uniformly as $T \to \infty$, for $|\omega| > \epsilon$.*

Kernels can be applied to both autocovariance and periodogram estimates. When applied to the periodogram, a kernel produces an estimate of the spectrum at frequency ω using a weighted average of the values of the periodogram in a neighborhood of ω. Note that this neighborhood is shrinking as $T \to \infty$, since the bias in ACF estimates asymptotically disappears. Hence, in the limit, $\mathcal{K}_T(\omega)$ looks like a δ-function, i.e. it puts all its mass at one point.

There are several types of kernels. Those used in this book are the following:

1) Box-Car (Truncated) $\mathcal{K}_{TR}(\omega) = \begin{cases} 1 & if\ |\omega| \leq J(T) \\ 0 & otherwise \end{cases}$

2) Bartlett $\mathcal{K}_{BT}(\omega) = \begin{cases} 1 - \frac{|\omega|}{J(T)} & if\ |\omega| \leq J(T) \\ 0 & otherwise \end{cases}$

3) Parzen $\mathcal{K}_{PR}(\omega) = \begin{cases} 1 - 6(\frac{\omega}{J(T)})^2 + 6(\frac{|\omega|}{J(T)})^3 & 0 \leq |\omega| \leq J(T)/2 \\ 2(1 - \frac{|\omega|}{J(T)})^3 & J(T)/2 \leq |\omega| \leq J(T) \\ 0 & otherwise \end{cases}$

4) Quadratic spectral $\mathcal{K}_{QS}(\omega) = \frac{25}{12\pi^2\omega^2}(\frac{\sin(6\pi\omega/5)}{6\pi\omega/5} - \cos(6\pi\omega/5))$

Here $J(T)$ is a truncation point, typically chosen to be a function of the sample size T. Note that the quadratic spectral kernel has no truncation point. However, it is useful to

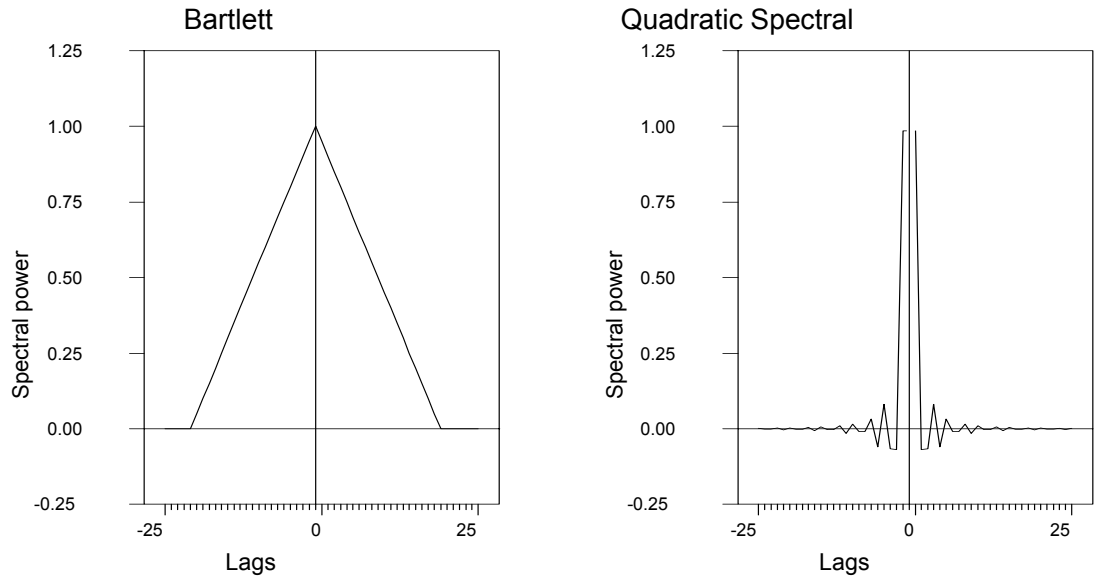


Figure 1.4: Kernels

define the first time that $\mathcal{K}_{QS}$ crosses zero (call it $J^*(T)$) and this point plays the same role as $J(T)$ in the other three kernels.

The Bartlett kernel and the quadratic spectral kernel are the most popular ones. The Bartlett kernel has the shape of a tent with width $2J(T)$. To insure consistency of the spectral estimates, it is standard to select $J(T)$ so that $\frac{J(T)}{T} \to 0$ as $T \to \infty$. In figure 1.4 we have set J(T)=20. The quadratic spectral kernel has the form of a wave with infinite loops, but after the first crossing, side loops are small.

Exercise 1.32 *Show that the coherence estimator* $\widehat{Co(\omega)} = \frac{|\widehat{\mathcal{S}}_{y_1,y_2}(\omega)|}{\sqrt{\widehat{\mathcal{S}}_{y_1,y_1}(\omega)\widehat{\mathcal{S}}_{y_2,y_2}(\omega)}}$ *is consistent, where* $\widehat{\mathcal{S}}_{y_i,y_{i'}}(\omega) = \frac{1}{2\pi}\sum_{\tau=-T+1}^{T-1}\widehat{ACF}_{y_i,y_{i'}}(\tau)\mathcal{K}_T(\omega)e^{-i\omega\tau}$, $\mathcal{K}_T(\omega)$ *is a kernel and* $i, i' = 1, 2$.

While for most part of this book we will consider stationary processes, we will deal at times with processes which are only locally stationary (e.g. processes with time varying coefficients). For these processes, the spectral density is not defined. However, it is possible to define a "local" spectral density and practically all the properties we have described apply also to this alternative construction. For details, see Priestley (1980, chapter 11).

Exercise 1.33 *Compute the spectral density of consumption, investment, output, hours, real wage, consumer prices, M1 and the nominal interest rate using quarterly US data and compute their pairwise coherence with output. Are there any interesting features at business cycle frequencies you would like to emphasize? Repeat the exercise using EU data. Are there important differences with the US? (Hint: Careful with potential non-stationarities in the data).*

Chapter 2: DSGE Models, Solutions and Approximations

This chapter describes some standard Dynamics Stochastic General Equilibrium (DSGE) models which will be used in examples and exercises throughout the book. Since these models do not have a closed form solution, except in very special circumstances, we also present a number of methods to obtain approximate solutions to the optimization problems.

There is a variety of models currently used in macroeconomics. The majority is based on two simple setups: a competitive structure, where allocations are, in general, Pareto optimal; and a monopolistic competitive structure, where one type of agents can set the price of the goods she supplies and allocations are suboptimal. Typically, an expression for the variables of interest in terms of the exogenous forces and the states is found in two ways. When competitive allocations are Pareto optimal the principle of dynamic programming is typically used and iteration on Bellman equation are employed to compute the value function and the policy rules, whenever they are known to exist and to be unique. As we will see, calculating the value function is a complicated enterprise except with simple but often economically unpalatable specifications. For general preference and technological specifications, quadratic approximations of the utility function, and discretization of the dynamic programming problem are generally employed.

When the equilibrium allocations are distorted, one must alter the dynamic programming formulation and in that case Bellman equation does not have an hedge over a more standard stochastic Lagrangian multipliers methodology, where one uses the first order conditions, the constraints and the transversality condition, to find a solution. Solutions are hard to find also with the Lagrangian approach since the problem is nonlinear and it involves expectations of future variables. Euler equation methods, which approximate the first order conditions, the expectational equations or the policy function can be used in these frameworks. Many methods have been developed in the literature. Here we restrict attention to the three widely used approaches: discretization of the state and shock space; log-linear and second order approximations; and parametrizing expectations. For a thorough discussion of the various methodologies see Cooley (1995, chapters 2 and 3) or Marimon and Scott (1999).

The next two sections illustrate features of various models and the mechanics of different solution methods with the aid of examples and exercises. A comparison between various

approaches concludes the chapter.

2.1 Few useful models

It is impossible to provide a thorough description of the models currently used in macroeconomics. Therefore, we focus attention on two prototype structures: one involving only real variables and one considering also nominal ones. In each case, we analyze models with both representative and heterogeneous agents and consider both optimal and distorted setups.

2.1.1 A basic Real Business Cycle (RBC) Model

A large portion of the current macroeconomic literature uses versions of the one sector growth model to jointly explain the cyclical and the long-run properties of the data. In the basic setup we consider there is a large number of identical households that live forever and are endowed with one unit of time, which they can allocate to leisure or to work, and K_0 unit of productive capital, which depreciates at the rate $0 < \delta < 1$ every period. The social planner chooses $\{c_t, N_t, K_{t+1}\}_{t=0}^{\infty}$ to maximize

$$\max E_0 \sum_t \beta^t u(c_t, c_{t-1}, N_t) \tag{2.1}$$

where c_t is consumption, N_t is employment (hours) and K_t is capital and $E_0 \equiv E[.|\mathcal{F}_0]$ is the expectations operator, conditional on the information set $\mathcal{F}_0$, $0 < \beta < 1$. The instantaneous utility function is bounded, twice continuously differentiable, strictly increasing and strictly concave in all arguments. It depends on c_t and c_{t-1} to account for possible habit formation in consumption. The maximization of (2.1) is subject to the sequence of constraints

$$\begin{aligned} c_t + K_{t+1} &\leq (1 - T^y) f(K_t, N_t, \zeta_t) + T_t + (1 - \delta) K_t && (2.2) \\ 0 \leq N_t &\leq 1 && (2.3) \end{aligned}$$

where $f(.)$ is twice continuously differentiable, strictly increasing and strictly concave in K_t and N_t production technology; ζ_t is a technological disturbance, T^y is a (constant) income tax rate and T_t lump sum transfers.

There is a government which finances a stochastic flow of current expenditure with income taxes and lump sum transfers: expenditure is unproductive and does not yield utility for the agents. We assume a period-by-period balanced budget of the form

$$G_t = T^y f(K_t, N_t, \zeta_t) - T_t \tag{2.4}$$

The economy is closed by the resource constraint, which provides a national account identity:

$$c_t + K_{t+1} - (1 - \delta) K_t + G_t = f(K_t, N_t, \zeta_t) \tag{2.5}$$

Note that in (2.4) we have assumed that the government balances the budget at each t. This is not restrictive since agents in this economy are Ricardian; that is, the addition of government debt does not change optimal allocations. This is because, if debt is held in equilibrium, it must bear the same rate of return as capital, so that $(1 + r_t^B) = E_t[f_k(1 - T^y) + (1 - \delta)]$, where $f_k = \frac{\partial f}{\partial K}$. In other words, debt is a redundant asset and can be priced by arbitrage, once (δ, T^y, f_k) are known. One example where debt matters is considered later on.

Exercise 2.1 *Decentralize the RBC model so that there is a representative consumer and a representative firm. Assume that the consumer makes the investment decision while the firm hires capital and labor from the consumer. Is it true that decentralized allocations are the same as those obtained in the social planner's problem? What conditions need to be satisfied? Repeat the exercise assuming that the firm makes the investment decision.*

Exercise 2.2 *Set $c_{t-1} = 0$ in (2.1) and assume $T^y = 0$ $\forall t$.*
i) Define the variables characterizing the state of the economy at each t (the states) and the choice variables (the controls).
ii) Show that the problem described by (2.1)-(2.5) can be equivalently written as:

$$\mathbb{V}(K, \zeta, G) = \max_{\{K^+, N\}} \{u([f(K, N, \zeta) + (1-\delta)K - G - K^+], N) + \beta E[\mathbb{V}(K^+, \zeta^+, G^+)|K, \zeta, G]\} \tag{2.6}$$

$0 < N_t < 1$ where the value function $\mathbb{V}$ is the utility value of the optimal plan, given (K_t, ζ_t, G_t), $E(\mathbb{V}|.)$ is the expectation of $\mathbb{V}$ conditional on the available information and the superscript $''+''$ indicates future values.
iii) Assume $u(c_t, c_{t-1}, N_t) = \ln c_t + \ln(1 - N_t)$ and that $GDP_t \equiv f(K_t, N_t, \zeta_t) = \zeta_t K_t^{1-\eta} N_t^{\eta}$. Find steady state values for $(\frac{K}{GDP}, \frac{c}{GDP}, N)$.

Note that (2.6) define the so-called Bellman equation, a functional equation giving the maximum value of the problem for each value of the states and the shocks, given that from next period on agent behave optimally.

There are few conditions that need to be satisfied for a model to be fitted into a Bellman equation format. First, the utility function must be time separable in the contemporaneous control and state variables. Second, the objective function and the constraints have to be such that current decisions affect current and future utilities but not past ones. While these conditions are typically satisfied, there are situations where Bellman equation (and its associated optimality principle) may fail to characterize particular economic problem. One is the time inconsistency problem analyzed by Kydland and Prescott (1977), a version of which is described in the next example.

Example 2.1 *Suppose agents in the economy maximize $E_0 \sum_t \beta_t (\ln c_t + \gamma \ln \frac{B_t}{p_t})$ subject to $c_t + B_t p_t \leq w_t + B_{t-1} p_t + T_t \equiv We_t$, where B_t are goverment backed assets, w_t is labor income and T_t lump sum taxes (tranfers), by choice of sequences for c_t and B_t, given T_t, p_t. The goverment budget constraint is $g_t = B_t p_t - B_{t-1} p_t + T_t$ where g_t is random.*

We assume that the goverment chooses B_t to maximize agents' welfare. Agents' problem is recursive. In fact, using their wealth We_t as a state for the problem, Bellman equation is $\mathbb{V}(We) = \max_{c,B}(\ln c + \gamma \ln \frac{B}{p}) + \beta \mathbb{V}(We^+)$ and the constraint is $We = c + \frac{B}{p}$. The first order conditions for the problem can be summarized via $\frac{1}{c_t p_t} = \beta \frac{1}{c_{t+1} p_{t+1}} + \frac{\gamma}{B_t}$. Therefore, solving forward and using the resource constraint, we have

$$\frac{1}{p_t} = \gamma(w_t - g_t) \sum_{j=0}^{\infty} \beta^j \frac{1}{B_{t+j+1}} \tag{2.7}$$

The goverment takes (2.7) as given and maximizes agents' utility subject to the resource constraint. Substituting (2.7) into the utility function we have

$$\max_{B_t} \sum_t \beta^t (\ln c_t + \gamma \ln(B_t \gamma (w_t - g_t) \sum_{j=0}^{\infty} \beta^j \frac{1}{B_{t+j+1}}) \tag{2.8}$$

Clearly in (2.8) future controls B_t affect current utility. Therefore, the goverment problem can not be cast into a Bellman equation.

Exercise 2.3 *Show how to modify Bellman equation (2.6) when $T^y \neq 0$.*

A solution to (2.6) is typically hard to find since $\mathbb{V}$ is unknown and there is no analytic expression for it. Had the solution been known, we could have used (2.6) to define a function h mapping every (K, G, ζ) into (K^+, N) that gives the maximum.

Since $\mathbb{V}$ is unknown, methods to prove its existence and uniqueness and to describe its properties have been developed (see e.g. Lucas and Stokey (1989)). These methods implicitly provide a way of computing a solution to (2.6) which we summarize next:

Algorithm 2.1

1) *Choose a differentiable and concave function $\mathbb{V}^0(K, \zeta, G)$.*

2) *Compute $\mathbb{V}^1(K, \zeta, G) = \max_{\{K^+, N\}} \{u([f(K, \zeta, N) + (1 - \delta)K - G - K^+], N) + \beta E[\mathbb{V}^0(K^+, \zeta^+, G^+)|K, \zeta, G]\}$.*

3) *Set $\mathbb{V}^0 = \mathbb{V}^1$ and iterate on 2) until $|\mathbb{V}^{l+1} - \mathbb{V}^l| < \iota$, ι small.*

4) *When $|\mathbb{V}^{l+1} - \mathbb{V}^l| < \iota$, compute $K^+ = h_1(K, \zeta, G)$ and $N = h_2(K, \zeta, G)$.*

Hence, $\mathbb{V}$ can be obtained as the limit of $\mathbb{V}^l$ for $l \to \infty$. Under regularity conditions, this limit exists, it is unique and the sequence of iterations defined by algorithm 2.1 achieves it.

For simple problems algorithm 2.1 is fast and accurate. For more complicated ones, when the combined number of states and shocks is large, it may computationally demanding. Moreover, unless $\mathbb{V}^0$ is appropriately chosen, the iteration process may be time consuming. In a few simple cases, the solution to Bellman equation has a known form and the simpler method of undetermined coefficients can be used. We analyze one of these cases in the next example.

Example 2.2 *Assume that* $u(c_t, c_{t-1}, N_t) = \ln c_t + \vartheta_n \ln(1 - N_t)$, $\delta = 1$; *the production function has the form* $GDP_{t+1} = \zeta_{t+1} K_t^{1-\eta} N_t^{\eta}$; *the resource constraint is* $GDP_t = K_t + c_t$, $\ln \zeta_t$ *is an AR(1) process with persistence* ρ, *and set* $G_t = T^y = T_t = 0$. *The states of the problem are* GDP_t *and* ζ_t *while the controls are* c_t, K_t, N_t. *We guess that the value function has the form* $\mathbb{V}(K, \zeta) = \mathbb{V}_0 + \mathbb{V}_1 \ln GDP_t + \mathbb{V}_2 \ln \zeta_t$. *Then Bellman equation maps logarithmic functions into logarithmic ones. Therefore the limit, if it exists, will also have a logarithmic form. To find the constants* $\mathbb{V}_0, \mathbb{V}_1, \mathbb{V}_2$ *we proceed as follows. First we substitute the constraint into the utility function and use the guess to solve out future GDP. That is:*

$$\begin{aligned} \mathbb{V}_0 + \mathbb{V}_1 \ln GDP_t + \mathbb{V}_2 \ln \zeta_t &= [\ln(GDP_t - K_t) + \vartheta_N (1 - N_t)] + \beta \mathbb{V}_0 + \beta \mathbb{V}_1 (1 - \eta) \ln K_t \\ &+ \beta \mathbb{V}_1 \eta \ln N_t + \beta E_t (\mathbb{V}_2 + \mathbb{V}_1) \ln \zeta_{t+1} \end{aligned} \tag{2.9}$$

Maximizing (2.9) with respect to (K_t, N_t) *we have* $N_t = \frac{\beta \mathbb{V}_1 \eta}{\vartheta_N + \beta \mathbb{V}_1 \eta}$ *and* $K_t = \frac{\beta(1-\eta)\mathbb{V}_1}{1+\beta(1-\eta)\mathbb{V}_1} GDP_t$. *Substituting these expressions into (2.9) and using the fact that* $E_t \zeta_{t+1} = \rho \zeta_t$ *we obtain*

$$\mathbb{V}_0 + \mathbb{V}_1 \ln GDP_t + \mathbb{V}_2 \ln \zeta_t = \text{constants} + (1 + (1-\eta)\beta \mathbb{V}_1) \ln GDP_t + \beta \rho (\mathbb{V}_2 + \mathbb{V}_1) \ln \zeta_t \tag{2.10}$$

Matching coefficients on the two sides of the equation we have $1 + (1 - \eta)\beta \mathbb{V}_1 = \mathbb{V}_1$ *or* $\mathbb{V}_1 = \frac{1}{1-(1-\eta)\beta}$ *and* $\beta\rho(\mathbb{V}_2 + \mathbb{V}_1) = \mathbb{V}_2$ *or* $\mathbb{V}_2 = \frac{\rho\beta}{(1-(1-\eta)\beta)^2}$. *Using the solution for* $\mathbb{V}_1$ *into the expressions for* K_t, N_t *we have that* $K_t = (1-\eta)\beta GDP_t$ *and* $N_t = \frac{\beta\eta}{\vartheta_N(1-\beta(1-\eta)+\beta\eta}$ *and from the resource constraints* $c_t = (1 - (1-\eta)\beta) GDP_t$. *Hence, with this preference specification, the optimal labor supply decision is very simple: keep hours constant, regardless of the state and the shocks of the economy.*

Exercise 2.4 *Assume that* $u(c_t, c_{t-1}, N_t) = \ln c_t, \delta = 1$; *the production function has the form* $GDP_t = \zeta_t K_t^{1-\eta} N_t^{\eta}$; *the resource constraint is* $c_t + K_{t+1} + G_t = GDP_t$, *and* $G_t = T_t$ *that both* (ζ_t, G_t) *are iid. Guess that the value function is* $\mathbb{V}(K, \zeta, G) = \mathbb{V}_0 + \mathbb{V}_1 \ln K_t + \mathbb{V}_2 \ln \zeta_t + \mathbb{V}_3 \ln G_t$. *Determine* $\mathbb{V}_0, \mathbb{V}_1, \mathbb{V}_2, \mathbb{V}_3$. *Show the optimal policy for* K^+.

Two other cases where a solution to the Bellman equation can be found analytically are analyzed in the next exercise.

Exercise 2.5 *i) Suppose that* $u(c_t, c_{t-1}, N_t) = a_0 + a_1 c_t - a_2 c_t^2$ *and that* $G_t = T_t = T^y = 0 \ \forall t$. *Show that the value function is of the form* $\mathbb{V}(K, \zeta) = [K, \zeta]' \mathbb{V}_2 [K, \zeta] + \mathbb{V}_0$. *Find the values of* $\mathbb{V}_0$ *and* $\mathbb{V}_2$. *(Hint: use the fact that* $E(e_t' \mathbb{V}_2 e_t) = tr(\mathbb{V}_2) E(e_t' e_t) = tr(\mathbb{V}_2)\sigma_e^2$ *where* σ_e^2 *is the covariance matrix of* e_t *and* $tr(\mathbb{V}_2)$ *is the trace of* $\mathbb{V}_2$*). Show that the decision rule for* c *and* K^+ *is linear in* K *and* ζ.
ii) Suppose $u(c_t, c_{t-1}, N_t) = \frac{c_t^{1-\varphi}}{1-\varphi}$; $K_t = 1 \ \forall t$ *and assume that* ζ_t *can take three values. Let* ζ_t *evolve according to* $P(\zeta_t = i | \zeta_{t-1} = i') = p_{ii'} > 0$. *Assume that there are claims to the output in the form of stocks* S_t, *with price* p_t^s *and dividend* sd_t. *Write down Bellman equation. Let* $\beta = 0.9$, $p_{ii} = 0.8$, $i = 1, \ldots, 3$; $p_{i,i+1} = 0.2$ *and* $p_{ii'} = 0$ *otherwise. Show the first two terms of the value function iterations. Can you guess what the limit is?*

We can relax some of the assumptions we have made (e.g. we can use a more general law of motion for the shocks), but except for these simple cases, even the most basic stochastic RBC model does not have a closed form solution. As we will see later, existence of a closed form solution is not necessary to estimate the structural parameters (here β, δ, η), and the parameters of the process for ζ_t and G_t and to examine its fit to the data. However, a solution is needed when one wishes to simulate the model, compare its dynamics with those of the data, and/or perform policy analyses.

There is an alternative to Bellman equation to solve simple optimization problems. It involves substituting all the constraints in the utility function and maximizing the resulting expression unconstrained or, if this is not possible, using a stochastic Lagrange multiplier approach. We illustrate the former approach next with an example.

Example 2.3 *Suppose a representative consumer obtains utility from the services of durable and nondurable goods according to* $E_0 \sum_t \beta^t (cs_t - \upsilon_t)'(cs_t - \upsilon_t)$, *where* $0 < \beta < 1$, υ_t *is a preference shock and consumption services* cs_t *satisfy* $cs_t = b_1 cd_{t-1} + b_2 c_t$, *where* cd_{t-1} *is the stock of durable goods, accumulated according to* $cd_t = b_3 cd_{t-1} + b_4 c_t$, *where* $0 < b_1, b_3 < 1$ *and* $0 < b_2, b_4 \leq 1$ *are parameters. Output is produced with the technology* $f(K_{t-1}, \zeta_t) = (1-\eta)K_{t-1} + \zeta_t$, *where* $0 < \eta \leq 1$ *and* ζ_t *is a productivity disturbance, and divided between consumption and investment goods according to* $b_5 c_t + b_6 inv_t = GDP_t$. *Physical capital accumulates according to* $K_t = b_7 K_{t-1} + b_8 inv_t$, *where* $0 < b_7 < 1,\ 0 < b_8 \leq 1$.

Using the definition of $(cs_t,\ cd_t,\ K_t)$ *and the resource constraint we have*

$$cs_t + cd_t = (b_1 + b_3) cd_{t-1} + \frac{(b_2 + b_4)}{b_5}((1-\eta)K_{t-1} + \zeta_t - \frac{b_6}{b_8}(K_t - b_7 K_{t-1})) \tag{2.11}$$

Letting $b_9 = (b_1 + b_3),\ b_{10} = \frac{(b_2+b_4)}{b_5},\ b_{11} = b_{10}\frac{b_6}{b_8}, b_{12} = b_{11} b_7$ *and using (2.11) in the utility function the problem can be reformulated as*

$$\max_{\{cd_t, K_t\}} E_0 \sum_t \beta^t \mathcal{C}_1[cd_t, K_t]' + \mathcal{C}_2[cd_{t-1}, k_{t-1}, \zeta_t, \upsilon_t]')'(\mathcal{C}_1[cd_t, K_t]' + \mathcal{C}_2[cd_{t-1}, k_{t-1}, \zeta_t, \upsilon_t]') \tag{2.12}$$

where $\mathcal{C}_1 = [-1, -b_{11}], \mathcal{C}_2 = [b_9, b_{12} + b_{10}(1-\eta), b_{10}, -1]$. *If* $\mathcal{C}_1'\mathcal{C}_1$ *is invertible, and the shocks* (ζ_t, υ_t) *are known at each* t, *the first order condition of the model imply* $[cd_t, K_t]' = (\mathcal{C}_1'\mathcal{C}_1)^{-1}(\mathcal{C}_1'\mathcal{C}_2)\ [cd_{t-1}, K_{t-1}, \zeta_t, \upsilon_t]'$. *Given* $(cd_t, K_t, \zeta_t, \upsilon_t)$, *values for* cs_t *and* c_t *can be found from (2.11) and from consumption services constraint.*

Economic models with quadratic objective functions and linear constraints can also be cast into standard optimal control problem formats, which allows calculation of the solution with simple and fast algorithms.

Exercise 2.6 *Take the model of example 2.3 but let* $\upsilon_t = 0$. *Cast it into an optimal linear regulator problem of the form* $\max E_0 \sum_t \beta^t [y_{2t} \mathcal{Q}_2 y_{2t}' + y_{1t} \mathcal{Q}_1 y_{1t}' + 2 y_{2t} \mathcal{Q}_3 y_{1t}']$ *subject to* $y_{2t+1} = \mathcal{Q}_4 y_{2t} + \mathcal{Q}_5 y_{1t} + \mathcal{Q}_6 y_{3t+1}$ *where* y_{3t} *is a vector of (serially correlated) shocks,* y_{2t} *a vector of states and* y_{1t} *a vector of controls. Show the form of* $\mathcal{Q}_i, i = 1, \ldots, 6$.

. A stochastic Lagrange multiplier approach works even when Bellman equation does not characterize the problem under consideration but requires a somewhat stronger set of assumptions to be applicable. Basically, we need that the objective function is strictly concave, differentiable and that the derivatives have finite expectations; that the constraints are convex, differentiable and that the derivatives have finite expectations; that the choice variables are adapted to the information set; that expected utility is bounded and converges to a limit as $T \to \infty$ and that there exists a sequence of multipliers λ_t such that at the optimum the Kuhn-Tucker conditions hold with probability 1 (see Sims (2002) for a formal statement of these requirements).

It is straightforward to check that these conditions are satisfied for the simple RBC model we have considered so far. Then, letting $f_N = \frac{\partial f}{\partial N}$, $U_{c,t} = \frac{\partial u(c_t, c_{t-1}, N_t)}{\partial c_t}$, $U_{N,t} = \frac{\partial u(c_t, c_{t-1}, N_t)}{\partial N_t}$, the Euler equation for capital accumulation in the basic RBC model is

$$E_t \beta \frac{U_{c,t+1}}{U_{c,t}} [(1 - T^y) f_k + (1 - \delta)] - 1 = 0 \tag{2.13}$$

while the intratemporal marginal condition between consumption and labor is:

$$\frac{U_{c,t}}{U_{N,t}} = -\frac{1}{(1 - T^y) f_N} \tag{2.14}$$

(2.13)-(2.14), the budget constraint and the transversality condition, $\lim_{t\to\infty} \sup \beta^t (U_{j,t} - \lambda_t g_{j,t})(j_t - \hat{j}_t) \leq 0$, where $j = c, N$,$g_{j,t}$ is the derivative of the constraints with respect to j, $\hat{j}_t$ is the optimal choice and j_t any other choice, then need to be solved for (K_{t+1}, N_t, c_t), given (G_t, ζ_t, K_t). This is not easy. Since the system of equations is nonlinear and involves expectations of future variables, no analytical solution exists in general.

Exercise 2.7 *Solve the problem of example 2.3 using a Lagrange multiplier approach. Show the conditions you need for the solution to be the same as in example 2.3.*

Versions of the basic RBC model with additional shocks, alternative inputs in the production function or different market structures have been extensively examined in the macroeconomic literature. We consider some of these extensions in the next four exercises.

Exercise 2.8 *(Utility producing government expenditure) Consider a basic RBC model and suppose that government expenditure provides utility to the agents; that private and public consumption are substitutes in the utility function; and that there is no habit in consumption, e.g.* $U(c_t, c_{t-1}, G_t, N_t) = (c_t + \vartheta_G G_t)^{\vartheta} (1 - N_t)^{1-\vartheta}$.
i) Using steady state relationships describe how private and public consumption are related. Is there some form of crowding out?
ii) In a cross section of steady states, is it true that countries which have higher level of government expenditure will also have lower levels of leisure (i.e. is it true that the income effect of distortionary taxation is higher when G is higher)?

Exercise 2.9 *(Production externalities) In a basic RBC model assume that output is produced with firm specific inputs and the aggregate capital stock, i.e.* $f(K_{it}, N_{it}, \zeta_t, K_t) = K_t^{\aleph} K_{it}^{1-\eta} N_{it}^{\eta} \zeta_t$, $\aleph > 0$ *and* $K_t = \int K_{it} di$.
i) Derive the first order conditions and discuss how to find optimal allocations.
ii) Is it appropriate to use Bellman equation to find a solution to this problem? What modifications do you need to introduce to the standard setup?

Exercise 2.10 *(Non-competitive labor markets) Assume that, in a basic RBC model, there are one-period labor contracts. The contracts set the real wage on the basis of the expected marginal product of labor. Once shocks are realized, and given the contractual real wage, the firm chooses employment to maximize its profits. Write down the contractual wage equation and the optimal labor decision rule by firms. Compare it with a traditional Phillips curve relationship where* $\ln N_t - E_{t-1}(\ln N_t) \propto \ln p_t - E_{t-1}(\ln p_t)$.

Exercise 2.11 *(Capacity utilization) Assume* $G_t = T_t = T^y = 0$*; that the production function depends on capital* (K_t) *and its utilization* (ku_t) *and it is of the form* $f(K_t, ku_t, N_t, \zeta_t) = \zeta_t (K_t ku_t)^{1-\eta} N_t^{\eta}$. *This production function allows firms to respond to shocks by varying utilization even when the stock of capital is fixed. Assume that capital depreciates in proportion to its use. In particular, assume* $\delta(ku_t) = \delta_0 + \delta_1 ku_t^{\delta_2}$, *where* δ_0, δ_1 *and* δ_2 *are parameters.*
i) Write down the optimality conditions of the firm's problem and Bellman equation.
ii) Show that, if capital depreciates instantaneously, the solution of this problem is identical to the one of a standard RBC model examined in part ii) of exercise 2.2.

Although it is common to proxy for technological disturbances with Solow residuals, such an approach is often criticized in the literature. The main reason is that such a proxy tends to overstate the variability of these shocks and may capture not only technology but also other sources of disturbances. The example below provides a case where this can occur.

Example 2.4 *Suppose that output is produced with part-time hours* (N^P) *and full-time hours* (N^F) *according to the technology* $GDP_t = \zeta_t K_t^{1-\eta} (N_t^F)^{\eta} + \zeta_t K_t^{1-\eta} (N_t^P)^{\eta}$. *Typically, Solow accounting proceeds assuming that part-time and full time hours are perfect substitutes and use total hours in the production function, i.e.* $GDP_t = \zeta_t K_t^{1-\eta} (N_t^F + N^P)^{\eta}$. *An estimate of* ζ_t *is obtained via* $\widehat{\ln \zeta_t} = \ln GDP_t - (1-\eta) \ln K_t - \eta \ln(N_t^F + N_t^P)$, *where* η *is the share of labor income. It is easy to see that* $\widehat{\ln \zeta_t} = \ln \zeta_t + \ln((N_t^F)^{\eta} + (N^P)^{\eta}) - \eta \ln(N_t^F + N^P)$, *so that the variance of* $\widehat{\ln \zeta_t}$ *overestimates the variance of* $\ln \zeta_t$. *This is a general problem: whenever a variable is omitted from an estimated equation, the variance of the estimated residuals is at least as large as the variance of the true one. Note also that if* $N_t^F > N_t^P$ *and if* N_t^F *is less elastic than* N_t^P *to shocks (e.g., if there are differential cost in adjusting full and part-time hours),* $\ln((N_t^F)^{\eta} + (N^P)^{\eta}) - \eta \ln(N_t^F + N^P) > 0$. *In this situation any (preference) shock which alters the relative composition of* N^F *and* N^P *could induce procyclical labor productivity movements, even if* $\zeta_t = 0, \ \forall t$.

Several examples in this book are concerned with the apparently puzzling correlation between hours (employment) and labor productivity - the so-called Dunlop-Tashis puzzle.

What is puzzling is that this correlation is roughly zero in the data while is high and positive in an RBC model. We will study later how demand shocks can affect the magnitude of this correlation. In the next example we examine how the presence of government capital alters this correlation when an alternative source of technological disturbance is considered.

Example 2.5 *(Finn) Suppose agents' utility is* $u(c_t, c_{t-1}, N_t) = \frac{(c_t^{\vartheta}(1-N_t)^{1-\vartheta})^{1-\varphi}}{1-\varphi}$*, the budget constraint is* $(1-T^y)w_tN_t + (r_t - T^K(r_t-\delta))K_t^P + T_t + (1+r_t^B)B_t = c_t + inv_t^P + B_{t+1}$ *and private capital evolves according to* $K_{t+1}^P = (1-\delta)K_t^P + inv_t^P$*, where* T^K *are capital taxes,* r^B *is the net rate on real bonds and* r_t *the net return on private capital. Suppose the Government budget constraint is* $T^y w_t N_t + T^K(r_t-\delta)K_t^P + B_{t+1} = inv_t^G + T_t + (1+r_t^B)B_t$*, and government investments increase government capital according to* $inv_t^G = K_{t+1}^G - (1-\delta)K_t^G$*. The production function is* $GDP_t = \zeta_t N^{\eta}(K^P)^{1-\eta}(K^G)^{\aleph}$ *and* $\aleph \geq 0$*. Output is used for private consumption and investment.*

This model does not have an analytic solution but some intuition on how hours and labor productivity move can be obtained analyzing the effects of random variations in inv_t^G*. Suppose that* inv_t^G *is higher than expected. Then, less income is available for private use and, at the same time, more public capital is available in the economy. Which will be the dominant factor depends on the size of the investment increase relative to* $\aleph$*. If it is small, there will be a positive instantaneous wealth effect so that hours, investment and output decline while consumption and labor productivity increases. If it is large, a negative wealth effect will result in which case hours and output will increase and consumption and labor productivity decreases. In both cases, despite the RBC structure, the contemporaneous correlation between hours and labor productivity will be negative.*

2.1.2 Heterogeneous agent models

Although representative agent models constitute the backbone of current dynamic macroeconomics, the literature has started examining setups where some heterogeneities in either preferences, the income process, or the type of constraints that agent face are allowed for. The presence of heterogeneities does not change the structure of the problem: it is only required that the sum of individual variables match aggregate ones and that the planner problem is appropriately defined. The solution still requires casting the problem into a Bellman equation or setting up a stochastic Lagrange multiplier structure.

We consider few of these models here. Since the scope is purely illustrative we restrict attention to situations where there are only two types of agents. The generalization to a larger but finite number of agents' type is straightforward.

Example 2.6 *(A two country model with full capital mobility) Consider two countries and one representative agent in each country. Agents in country i choose sequences for consumption, hours, capital and contingent claim holdings to maximize* $E_0 \sum_{t=0}^{\infty} \beta^t \frac{[c_{it}^{\vartheta}(1-N_{it})^{1-\vartheta}]^{1-\varphi}}{1-\varphi}$ *subject to the following constraint*

$$c_{it} + \sum_j B_{jt+1} p_{jt}^B \leq B_{jt} + w_{it}N_{it} + r_{it}K_{it} - (K_{it+1} - (1-\delta)K_{it} - \frac{b}{2}(\frac{K_{it+1}}{K_{it}} - 1)^2)K_{it} \quad (2.15)$$

where $w_{it}N_{it}$ is labor income; $r_{it}K_{it}$ is capital income; B_{jt} is a set of Arrow-Debreu one period contingent claims and p^B_{jt} is its price; b is an adjustment cost parameter and δ the depreciation rate of capital. Since financial markets are complete, agents can insure themselves against all form of idiosyncratic risk.

We assume that factors of production are immobile. Domestic consumers rent capital and labor to domestic firms which produce an homogeneous intermediate good using a constant returns to scale technology. Domestic markets for factors of production are competitive and intermediate firms maximize profits. Intermediate goods are sold to domestic and foreign final good producing firms. The resource constraints are:

$$inty^1_{1t} + inty^1_{2t} = \zeta_{1t} K^{1-\eta}_{1t} N^{\eta}_{1t} \tag{2.16}$$

$$inty^2_{1t} + inty^2_{2t} = \zeta_{2t} K^{1-\eta}_{2t} N^{\eta}_{2t} \tag{2.17}$$

where $inty^1_{2t}$ are exports of goods from country 1 and $inty^2_{1t}$ imports from country 2.

Final goods are an aggregate of the goods produced by intermediate firms of the two countries. They are assembled with a constant returns to scale technology $GDP_{it} = (a_i(inty^1_{it})^{1-a_3}) + (1-a_i)(inty^2_{it})^{1-a_3}))^{\frac{1}{1-a_3}}$, where $a_3 \geq -1$ while a_1 and $(1-a_2)$ measure the domestic content of domestic spending. The resource constraint in the final goods market is $GDP_{it} = c_{it} + inv_{it}$. The two countries differ in the realizations of technology shocks. We assume $\ln(\zeta_{it})$ is an AR(1) with persistence $|\rho_\zeta| < 1$ and variance σ^2_ζ.

To map this setup into a Bellman equation assume that there is a social planner who attributes the weights $\mathbb{W}_1$ and $\mathbb{W}_2$ to the utilities of the agents of the two countries. Let the planner's objective function be $u^{sp}(c_{1t}, c_{2t}, N_{1t}, N_{2t}) = \sum_{i=1}^2 \mathbb{W}_i \, E_0 \sum_{t=0}^{\infty} \beta^t \frac{[c^\vartheta_{it}(1-N_{it})^{1-\vartheta}]^{1-\varphi}}{1-\varphi}$; let $y_{1t} = [inty^1_{it},\ inty^2_{it}, c_{it},\ N_{it},\ K_{it+1},\ B_{1t+1}, i = 1, 2]$, $y_{2t} = [K_{1t}, K_{2t}, B_{1t}]$ and $y_{3t} = [\zeta_{1t}, \zeta_{2t}]$. Then Bellman equation is $\mathbb{V}(y_2, y_3) = max_{\{y_1\}} u^{sp}(c_1, c_2, N_1, N_2) + E\beta\mathbb{V}(y^+_2, y^+_3 | y_2, y_3)$ and the constraints are given by (2.15) -(2.17) and the law of motion of the shocks.

Clearly, the value function has the same format as in (2.6). Since the functional form for utility is the same in both countries, the utility function of the social planner will also have the same functional form. Some information about the properties of the model can be obtained without computing a solution by examining the first order conditions and the properties of the final good production function. In fact, we have:

$$c_{it} + inv_{it} = p_{1t} inty^1_{it} + p_{2t} inty^2_{it} \tag{2.18}$$

$$Tot_t = \frac{p_{2t}}{p_{1t}} \tag{2.19}$$

$$nx_t = inty^1_{2t} - Tot_t inty^2_{1t} \tag{2.20}$$

(2.18) implies that output of the final good is allocated to the inputs according to their prices, $p_{2t} = \frac{\partial GDP_{1t}}{\partial inty^2_{1t}}$, $p_{1t} = \frac{\partial GDP_{1t}}{\partial inty^1_{1t}}$; (2.19) gives an expression for the terms of trade and (2.20) defines the trade balance.

Exercise 2.12 *Show that the demand functions for the two goods in country one are*

$$inty^1_{1t} = a_1^{\frac{1}{a_3}} (a_1^{\frac{1}{a_3}} + (1-a_1)^{\frac{1}{a_3}} Tot_t^{-\frac{1-a_3}{a_3}})^{-\frac{a_3}{1-a_3}} GDP_{1t}$$

$$inty_{1t}^2 = a_2^{\frac{1}{a_3}} Tot_t^{-\frac{1}{a_3}} (a_1^{\frac{1}{a_3}} + (1-a_1)^{\frac{1}{a_3}} Tot_t^{\frac{-(1-a_3)}{a_3}})^{-\frac{a_3}{1-a_3}} GDP_{1t}$$

i) Describe terms of trade relate to the variability of final goods demands.
ii) Noting that $Tot_t = \frac{(1-a_1)(inty_{1t}^2)^{-a_3}}{a_1(inty_{1t}^1)^{-a_3}}$ *show that when the elasticity of substitution between domestic and foreign good* $\frac{1}{a_3}$ *is high, any excess of demand in either of the two goods induces small changes in the terms of trade and large changes in the quantities used.*

Exercise 2.13 *Consider the same two country model of example 2.6 but now assume that financial markets are incomplete. That is, agents are forced to trade only a one-period bond which is assumed to be in zero net supply (i.e* $B_{1t} + B_{2t} = 0$*). How would you solve this problem? What does the assumption of incompleteness imply? Would it make a difference if agents of country 1 have limited borrowing capabilities, e.g.* $B_{1t} \leq K_{1t}$*?*

Interesting insights can be added to a basic RBC model when some agents are not optimizers.

Example 2.7 *Suppose that the economy is populated by standard RBC agents (their fraction in the total population is* Ψ*) which maximize* $E_0 \sum_t \beta^t u(c_t, c_{t-1}, N_t)$ *subject to the budget constraint* $c_t + inv_t + B_{t+1} = w_t N_t + r_t K_t + (1 + r_t^B) B_t + prf_t + T_t$*, where* prf_t *are firm's profits,* T_t *are government transfers and* B_t *are real bonds. Suppose that capital accumulates at the rate* $K_{t+1} = (1-\delta)K_t + inv_t$*. The remaining* $1 - \Psi$ *agents are myopic and consume all their income every period, that is* $c_t^{RT} = w_t N_t + T_t^{RT}$ *and supply all their work time inelastically at each* t*.*

Rule-of-thumb consumers play the role of an insensitive buffer in this economy. Therefore total hours, aggregate output and aggregate consumption will be much less sensitive to shocks than in an economy where all agents are optimizers. For example, goverment expenditure shocks crowd out consumption less and under some efficiency wage specification, they can even make it increase.

Exercise 2.14 *(Kiyotaki and Moore) Consider a model with two goods, land La, which is in fixed supply, and fruit which is non-storable, and a continuum of two types of agents: farmers of measure 1 and gatherers of measure* Ψ*. Both types of agents have utilities of the form* $E_t \sum_t \beta_j c_{j,t}$*, where* $c_{j,t}$ *is the consumption of fruit of agents type* j *and where* $\beta_{farmers} < \beta_{gatherers}$*. Let* p_t^L *be the price of land in terms of fruit and* r_t *the rate of exchange of a unit of fruit today for tomorrow. Both agents have technologies to produce fruit from land. Farmers use* $f(La_t)_{farmer} = (b_1 + b_2) La_{t-1}$ *where* b_1 *is the tradeable part and* b_2 *the bruised one (non-tradable), gatherers use* $f(La_t)_{gatherer}$*, where* $f_{gatherer}$ *is a decreasing returns to scale function and all output is tradable. The budget constraint for the two agents is* $p_t^L(La_{jt} - La_{jt-1}) + r_t B_{jt-1} + c_{jt}^\dagger = f(La_t) + B_{jt}$*, where* $c_{jt}^\dagger = c_{jt} + b_2 La_{t-1}$ *for farmers and* $c_{jt}^\dagger = c_{jt}$ *for gatherers,* B_{jt} *are loans and* $p_t^L(La_{jt} - La_{jt-1})$ *is the value of new land acquisitions. The farmers' technology is idiosyncratic so that only farmer* i *has the skill to produce fruit from it. Gatherer's technology does not require specific skills. Note*

that if no labor is used, fruit output is zero. i) Show that in equilibrium $r_t = r = \frac{1}{\beta_{gatherers}}$ *and that for farmers to be able to borrow a collateral is required. Show that the maximum amount of borrowing is* $B_t \leq \frac{p^L_{t+1} La_t}{r}$.
ii) Show that if there is no aggregate uncertainty, farmers borrow from gatherers up to their maximum, invest in land and consume $b_2 La_{t-1}$. *That is, for farmers* $La_t = \frac{1}{p^L_t - r^{-1} p^L_{t+1}}(b_1 + p^L_t) La_{t-1} - r B_{t-1}$ *where* $p^L_t - r^{-1} p^L_{t+1}$ *is the user cost of land (the down payment needed to purchase land) and* $B_t = r^{-1} p^L_{t+1} La_t$. *Argue that if* p^L_t *increases,* La_t *will increase (provided* $b_1 + p^L_t > r B_{t-1}/La_{t-1}$*) and* B_t *will also increase. Hence, the higher is the land price, the higher is the net worth of farmers and the more they will borrow.*

2.1.3 Monetary Models

The next set of models explicitly includes monetary factors. Finding a role for money in a general equilibrium model is difficult: with a full set of Arrow-Debreu claims, money is a redundant asset. Therefore, frictions of some sort need to be introduced for money to play some role. This means that the allocations produced by the competitive equilibrium are no longer optimal and that Bellman equation needs to be modified to include aggregate constraints. We focus attention on two popular specifications - a competitive model with transactional frictions and a monopolistic competitive framework where either sticky prices or sticky wages or both are exogenously imposed - and examine what they have to say about two questions: do monetary shocks generate liquidity effects? That is, do monetary shocks imply negative comovements between interest rates and money? Do expansionary monetary shocks imply expansionary and persistent output effects?

Example 2.8 *(Cooley and Hansen) The representative household maximizes* $E_0 \sum_t \beta^t u(c_{1t}, c_{2t}, N_t)$, *where* c_{1t} *is consumption of a cash good,* c_{2t} *is consumption of a credit good and* N_t *is hours worked. The budget constraint is:* $c_{1t} + c_{2t} + inv_t + \frac{M_{t+1}}{p_t} \leq w_t N_t + r_t k_t + \frac{M_t}{p_t} + \frac{T_t}{p_t}$ *where* $T_t = M_{t+1} - M_t$ *and* p_t *is the price level. There is a cash-in-advance constraint that forces households to buy* c_{1t} *with cash. We require* $p_t c_{1t} \leq M_t + T_t$ *and assume that the monetary authority sets* $\ln M^s_{t+1} = \ln M^s_t + \ln M^g_t$, *where* $\ln M^g_t$ *is an AR(1) process with mean* $\bar{M}$, *persistence* ρ_M *and variance* σ^2_M. *Households choose sequences for the two consumption goods, for employment, for investment and real balances to satisfy the budget constraint. We assume that shocks are realized at the beginning of each t so agents know their values when they take decisions. The resource constraint is* $c_{1t} + c_{2t} + inv_t = f(K_t, N_t, \zeta_t)$, *where* $\ln \zeta_t$ *is an AR(1) process with persistence* ρ_ζ *and variance* σ^2_ζ. *Since money is dominated in expected rate of return (by physical investments) the cash-in-advance constraint will be binding and agents hold just the exact amount of money needed to purchase* c_{1t}

When $\bar{M} > 0$, *money (and prices) grow over time. To map this setup into a stationary problem define* $M^*_t = \frac{M_t}{M^s_t}$ *and* $p^*_t = \frac{p_t}{M^s_{t+1}}$. *Then the value function is:*

$$V(K, k, M^*, \zeta, M^g) = U(\frac{M^* + \bar{M} - 1}{p^* \bar{M}} + \frac{T_t}{p_t}, wN + [r + (1-\delta)]k - k^+ - c_2 - \frac{(M^*)^+}{p^*}, N)$$

$$+ \quad \beta EV(K^+, k^+, (M^*)^+, \zeta^+, (M^g)^+) \tag{2.21}$$

where $K^+ = (1-\delta)K + Inv$, $k^+ = (1-\delta)k + inv$, $c_1 = \frac{M^* + \bar{M} - 1}{Mp^*} + \frac{T_t}{p_t}$ *and* K *represents the aggregate capital stock. The problem is completed by the consistency conditions* $K^+ = h_1(K, \zeta, M^g)$, $N = h_2(K, \zeta, M^g)$, $p^* = h_3(K, \zeta, M^g)$.

Not much can be done with this model without taking some approximation. However, we can show that monetary disturbances have perverse output effects and produce expected inflation but not liquidity effects. Suppose $c_{2t} = 0$, $\forall t$. *Then an unexpected increase in* M_t^g *makes agents substitute away from* c_{1t} *(which is now more expensive) toward credit goods - leisure and investment - which are cheaper. Hence, consumption and hours fall while investment increases. With a standard Cobb-Douglas production function output then declines. Also, since positive* M_t^g *shocks increase inflation, the nominal interest rate will increase, because both the real rate and expected inflation have temporarily increased. Hence, a surprise increase in* M_t^g *does not produce a liquidity effect nor output expansions.*

There are several ways to correct for the last shortcoming. For example, introducing one period labor contracts (as we have done in exercise 2.10) does change the response of output to monetary shocks. The next exercise provides a way to generate the right output and interest rates effects by introducing a loan market, forcing consumers to take decisions before shocks are realized and firms to borrow to finance their wage bill.

Exercise 2.15 *(Working capital) Consider the same economy of example 2.8 with* $c_{2t} = 0$ $\forall t$ *but assume that consumers deposit part of their money balances at the beginning of each* t *in banks. Assume that deposit decisions are taken before shocks occur and that firms face a working capital constraint, i.e. they have to pay for the factors of production before the receipts from the sale of the goods are received. Consumers maximize utility by choice of consumption, labor, capital and deposits, i.e.* $\max_{\{c_t, N_t, K_{t+1}, dep_t\}} E_0 \sum_t \beta^t \frac{(c_t^\vartheta (1-N_t)^{1-\vartheta})^{1-\varphi}}{1-\varphi}$. *There are three constraints. First, goods must be purchased with money, i.e.* $c_t p_t \leq M_t - dep_t + w_t N_t$. *Second, there is a budget constraint* $M_{t+1} = prf_{1t} + prf_{2t} + r_t p_t K_t + M_t - dep_t + w_t N_t - c_t p_t - inv_t p_t$, *where* $prf_{1t}(prf_{2t})$ *represent the share of firms' (banks') profits and* r_t *is the real return to capital. Third, capital accumulation is subject to an adjustment cost* $b \geq 0$, *i.e.* $inv_t = K_{t+1} - (1-\delta)K_t - \frac{b}{2}(\frac{K_{t+1}}{K_t} - 1)^2 K_t$. *Firms rent capital and labor and borrow cash from the banks to pay for the wage bill. Their problem is* $\max_{\{K_t, N_t\}} prf_{1t} = (p_t \zeta_t K_t^{1-\eta} N_t^\eta - p_t r_t K_t - (1+i_t) w_t N_t)$, *where* i_t *is the nominal interest rate. Banks take deposits and lend them together with new money to firms. Profits* prf_{2t} *are distributed, pro-rata to the households. The monetary authority sets its instrument according to*

$$M_t^{a_0} = i_t^{a_1} Y_t^{a_2} \pi_t^{a_3} M_t^g \tag{2.22}$$

where a_i *are parameters. For example, if* $a_0 = 0, a_1 = 1$, *the monetary authority sets the nominal interest rate as a function of output and inflation and stands ready to provide money when the economy needs it. Let* $(\ln \zeta_t, \ln M_t^g)$ *be AR(1) with persistence* ρ_ζ, ρ_M *and*

variances $\sigma_\zeta^2, \sigma_M^2$.
i) Set $b = 0$. *Show that the labor demand and the labor supply are:*

$$-U_{N,t} = \frac{w_t}{p_t} E_t \beta U_{c,t+1} \frac{p_t}{p_{t+1}}$$
$$\frac{w_t i_t}{p_t} = f_{N,t} \tag{2.23}$$

Argue that labor supply changes in anticipation of inflation while labor demand is directly affected by interest rates changes so that output will be positively related to money shocks.
ii) Show that the optimal saving decision satisfies $E_{t-1}\frac{U_{c,t}}{p_t} = E_{t-1} i_t \beta \frac{U_{c,t+1}}{p_{t+1}}$. *How does this compare with the saving decisions of the basic CIA model of example 2.8?*
iii) Show that the money demand can be written as $\frac{p_t GDP_t}{M_t} = \frac{1}{1+(\eta/i_t)}$, *where* $GDP_t = \zeta_t K_t^{1-\eta} N_t^\eta$. *Conclude that velocity* $\frac{p_t GDP_t}{M_t}$ *and the nominal rate are positively related and that a liquidity effect is generated in response to monetary disturbances.*

Exercise 2.16 *(Dunlop-Tarshis puzzle) Suppose agents maximize* $E_0 \sum_{t=0}^{\infty} \beta^t [\ln c_t + \vartheta_m \ln \frac{M_{t+1}}{p_t} + \vartheta_N \ln(1 - N_t)]$ *subject to* $c_t + \frac{M_{t+1}}{p_t} + K_{t+1} = w_t N_t + r_t K_t + (\frac{M_t + T_t}{p_t})$. *Let* $\pi_{t+1} = \frac{p_{t+1}}{p_t}$ *be the inflation rate. Firms rent capital from the households and produce using* $GDP_t = \zeta_t K_t^{1-\eta} N_t^\eta$, *where* $\ln \zeta_t$ *is a technological disturbance and capital depreciates in one period. Let the quantity of money evolve according to* $\ln M_{t+1}^s = \ln M_t^s + \ln M_t^g$ *and assume at each* t *the government takes away* G_t *units of output.*
i) Assume $G_t = G \ \forall t$. *Write down the first order conditions for the optimization problem of consumers and firms and find the competitive equilibrium for* $(c_t, K_{t+1}, N_t, w_t, r_t, \frac{M_{t+1}}{p_t})$.
ii) Show that, in equilibrium, employment is independent of the shocks, that output and employment are uncorrelated and that real wages are perfectly correlated with output.
iii) Show that monetary disturbances are neutral. Are they also superneutral (i.e. do changes in the growth rate of money have real effects)?
iv) Suppose there are labor contracts where the nominal wage rate is fixed one period in advance according to $w_t = E_{t-1} M_t + \ln(\eta) - \ln(\vartheta_m(\eta\beta)/(1-\beta)) - E_{t-1} \ln N_t$ *Show that monetary disturbances produce contemporaneous negative correlation between real wages and output.*
v) Now assume that G_t *is stochastic and set* $\ln M_t^g = 0 \ \forall t$. *What is the effect of government expenditure shocks on the correlation between real wages and output? Give some intuition for why adding labor contracts (Benassy) or government expenditure (Christiano and Eichenbaum) could reduce the correlation between real wages and output found in ii).*

The final type of model we consider adds nominal rigidities to a structure where monopolistic competitive firms produce intermediate goods which they sell to competitive final goods producers.

Example 2.9 *(Sticky prices) Suppose consumers choose* $\{c_t, N_t, K_{t+1}, M_{t+1}\}$ *to maximize* $E_0 \sum_t \beta^t (\frac{c_t^\vartheta (1-N_t)^{1-\vartheta})^{1-\varphi}}{1-\varphi} + \frac{1}{1-\varphi_m}(\frac{M_{t+1}}{p_t})^{1-\varphi_m})$ *subject to the budget constraint* $p_t(c_t + inv_t) + B_{t+1} + M_{t+1} \leq r_t p_t K_t + M_t + (1 + i_t) B_t + w_t N_t + prf_t$ *and the capital accumulation equation*

$inv_t = K_{t+1} - (1-\delta)K_t - \frac{b}{2}(\frac{K_{t+1}}{K_t} - 1))^2 K_t$, *where* b *is an adjustment cost parameter. Here* $prf_t = \int prf_{it}di$ *are profits obtained from owning intermediate firms. There are two types of firms: monopolistic competitive, intermediate good producing firms and perfectly competitive, final good producing firms. Final goods firms take the continuum of intermediate goods and bundle it up for final consumption. The production function for final goods is* $GDP_t = (\int_0^1 inty_{it}^{\frac{1}{1+\varsigma_p}} di)^{1+\varsigma_p}$, *where* ς_p *is the elasticity of substitution between intermediate goods. Profit maximization implies a demand for each input* i $\frac{inty_{it}}{GDP_t} = (\frac{p_{it}}{p_t})^{-\frac{1+\varsigma_p}{\varsigma_p}}$, *where* p_{it} *is the price of intermediate good* i *and* p_t *the price of the final good,* $p_t = (\int_0^1 p_{it}^{-\frac{1}{\varsigma_p}} di)^{-\varsigma_p}$.

Intermediate firms minimize costs and choose prices to maximize profits. Price decisions can not be taken every period: only $(1-\zeta_p)$ *of the firms are allowed to change prices at* t. *Their costs minimization problem is* $\min_{\{K_{it},N_{it}\}}(r_t K_{it} + w_t N_{it})$ *subject to* $inty_{it} = \zeta_t K_{it}^{1-\eta} N_{it}^{\eta}$ *and their profit maximization problem is* $\max_{\{p_{it+j}\}} \sum_j \frac{U_{c,t+j}}{p_{t+j}} \zeta_p^j prf_{it+j}$, *where* $\frac{U_{c,t+1}}{p_{t+1}}$ *is the value of a unit of profit,* prf_{it}, *to shareholders next period, subject to the demand function from final goods firms. Here* $prf_{t+j} = (p_{it+j} - mc_{it+j}) inty_{it+j}$ *and* $mc_{it} = \frac{w_{it}N_{it} + r_{it}K_{it}}{inty_{it}}$ *are marginal costs.*

We assume that the monetary authority uses a rule of the form (2.22). Since only a fraction of the firms can change prices at each t, *aggregate prices evolve according to:* $p_t = (\zeta_p p_{t-1}^{-\frac{1}{\varsigma_p}} + (1-\zeta_p)\tilde{p}_t^{-\frac{1}{\varsigma_p}})^{-\varsigma_p}$, *where* $\tilde{p}_t$ *is the common solution (all firms allowed to change prices are identical) to the following optimality condition (dropping the* i *subscript)*

$$0 = E_t \sum_j \beta^j \zeta_p^j \frac{U_{c,t+j}}{p_{t+j}} [\frac{\pi^j p_t}{1+\varsigma_p} - mc_{t+j}] inty_{t+j} \tag{2.24}$$

where π *is the steady state inflation rate. Hence, firms choose prices so that the discounted marginal revenues equals the discounted marginal costs in expected terms. Note that if* $\zeta_p \to 0$ *and no capital is present, (2.24) reduces to the standard condition that the real wage equals the marginal product of labor. (2.24) is the basis for the so-called New Keynesian Phillips curve (see e.g. Woodford (2003, ch. 3)), an expression relating current inflation to expected future inflation and to current marginal costs . To explicitly obtain such a relationship, (2.24) needs to be log-linearized around the steady state.*

To see what expression (2.24) involves, consider the case in which utility is logarithmic in consumption, linear in leisure and the marginal utility of real balances is negligible, i.e. $U(c_t, N_t, \frac{M_{t+1}}{p_t}) = \ln c_t + (1 - N_t)$, *output is produced with labor, prices are set every two periods and in each period half of the firms change their price. Optimal price setting is*

$$\frac{\tilde{p}_t}{p_t} = (1+\varsigma_p)(\frac{U_{c,t}c_t w_t + \beta U_{c,t+1}c_{t+1}w_{t+1}\pi_{t+1}^{\frac{1+\varsigma_p}{\varsigma_p}}}{U_{c,t}c_t + \beta U_{c,t+1}c_{t+1}\pi_{t+1}^{\frac{1}{\varsigma_p}}}) \tag{2.25}$$

where $\tilde{p}_t$ *is the optimal price,* p_t *the aggregate price level,* w_t *the wage rate and* $\pi_t = \frac{p_{t+1}}{p_t}$ *the inflation rate. Hence, ideally firms would like to charge a price which is a constant*

markup $(1+\varsigma_p)$ *over marginal (labor) costs. However, because the price level may change and prices are set for two periods, firms can't do this and chose a higher markup than* $(1+\varsigma_p)$ *in the period where prices are allowed to be changed. Note that if there are no shocks,* $\pi_{t+1}=1, w_{t+1}=w_t, c_{t+1}=c_t$ *and* $\frac{\tilde{p}_t}{p_t}=(1+\varsigma_p)w_t$.

Exercise 2.17 *i) Cast the sticky price model of example 2.9 into a Bellman equation formulation. Define states, controls and the value function.*
ii) Assuming that firms of type i have weight $\mathbb{W}_i$ *in the total, write down the first order conditions of the problem and interpret them.*
iii) Show that if all firms set prices one period in advance, the solution to (2.24) is $p_{it}=(1+\varsigma_p)E_{t-1}\frac{E_t\frac{U_{c,t+j}}{p_{t+j}}p_t^{\frac{1+\varsigma_p}{\varsigma_p}}inty_{it}}{E_{t-1}\frac{U_{c,t+j}}{p_{t+j}}p_t^{\frac{1+\varsigma_p}{\varsigma_p}}inty_{it}}mc_{it}$. *Conclude that if expectations do not change, firms set the price of goods as a constant markup over marginal costs.*
(iv) Intuitively explain why money expansions are likely to produce positive output effects. What conditions need to be satisfied for monetary expansions to produce a liquidity effect?

Extensions of the model that allow also for sticky wages are straightforward. We ask the reader to study a model with both sticky prices and sticky wages in the next exercise.

Exercise 2.18 *(Sticky wages) Assume that households are monopolistic competitive in the labor market so that they can choose the wage at which to work. Suppose capital is in fixed supply and that the period utility function is* $u_1(c_t)+u_2(N_t)+\frac{1}{1-\varphi_m}(\frac{M_{t+1}}{p_t})^{1-\varphi_m}$. *Suppose that households set nominal wages in a staggered way and that a fraction* $1-\zeta_w$ *can do this every period. When the household is allowed to reset the wage, she maximizes the discounted sum of utilities subject to the budget constraint.*
i) Show that utility maximization leads to:

$$E_t\sum_{j=0}^{\infty}\beta^j\zeta_w^j(\frac{\pi^j w_t}{(1+\varsigma_w)p_{t+j}}U_{1,t+j}+U_{2,t+j})N_{t+j}=0 \tag{2.26}$$

where β *is the discount factor, and* ς_w *is the elasticity of substitution in the labor aggregator* $N_t=(\int N_t(i)^{1/(1+\varsigma_w)}di)^{1+\varsigma_w}$, $i\in[0,1]$. *(Note: whenever the wage rate can not be changed* $w_{t+j}=\pi^j w_t$, *where* π *is the steady state inflation).*
ii) Show that if $\zeta_w=0$, *(2.26) reduces to* $\frac{w_t}{p_t}=-\frac{U_{2,t}}{U_{1,t}}$.
iii) Calculate equilibrium output, real rate and real wage when prices and wages are flexible.

Exercise 2.19 *(Taylor contracts, Edge) Consider a sticky wage model with no capital so that labor demand is* $N_t=y_t$, *real marginal costs are* $mc_t=w_t=1$ *and* $y_t=c_t$. *Suppose consumption and real balances are non substitutable in utility so that the money demand function is* $\frac{M_{t+1}}{p_t}=c_t$. *Suppose* $\ln M^s_{t+1}=\ln M^s_t+\ln M^g_t$, *where* $\ln M^g$ *is iid with mean* $\bar{M}>0$ *and assume two period staggered labor contracts.*

i) Show that the real wage satisfies $w_t = (0.5(\frac{\tilde{w}_t}{p_t})^{-\frac{1}{\varsigma_w}} + \frac{w_{t-1}}{p_t})^{-\frac{1}{\varsigma_w}})^{-\varsigma_w}$, *where* $\tilde{w}_t$ *is the nominal wage reset at* t.
ii) Show that $\pi_t = \frac{p_t}{p_{t-1}} = ((\frac{\tilde{w}_{t-1}}{p_{t-1}})^{-\frac{1}{\varsigma_w}}/(2-(\frac{\tilde{w}_t}{p_t})^{-\frac{1}{\varsigma_w}}))^{-\varsigma_w}$ *and that* $N_{it} = N_t((\frac{\tilde{w}_t}{p_t})/w_t)^{-\frac{1}{\varsigma_w}-1}$ *if the wage was set at* t *and* $N_{it} = N_t((\frac{\tilde{w}_{t-1}}{p_{t-1}})/(w_t\pi_t)^{-\frac{1}{\varsigma_w}-1}$ *if the wage was set at* $t-1$.
iii) Show that if utility is linear in N_t, *monetary shocks have no persistence.*

While expansionary monetary shocks produce expansionary output effects, their size is typically small and their persistence minimal, unless price stickiness is extreme. The next example shows a way to make output effects of monetary shocks sizable.

Example 2.10 *(Benhabib and Farmer) Consider an economy where output is produced with labor and real balances i.e.* $GDP_t = (a_1 N_t^{\eta} + a_2(\frac{M_t}{p_t})^{\eta})^{\frac{1}{\eta}}$. *Suppose agents' utility is* $u(c_t, n_t, N_t) = E_0 \sum_t \beta^t (\frac{c^{1-\varphi_c}}{1-\varphi_c} - \frac{1}{1-\varphi_n}\frac{n_t^{(1-\varphi_n)}}{N_t^{(\varphi_N-\varphi_n)}})$, *where* n_t *is individual employment,* N_t *is aggregate employment and* $\varphi_c, \varphi_n, \varphi_N$ *are parameters. The consumers' budget constraint is* $\frac{M_t}{p_t} = \frac{M_{t-1}}{p_t} + f(N_t, \frac{M_{t-1}+M_t^g}{p_t}) - c_t$ *and assume that* M_t^g *is iid with mean* $\bar{M} \geq 0$. *Equilibrium in the labor market implies* $\frac{-U_N}{U_c} = f_N(N_t, \frac{M_t}{p_t})$ *and the demand for money is* $E_t[\frac{\beta f_{M,t+1} U_{c,t+1}}{\pi_{t+1}}] = E_t[\frac{i_{t+1} U_{c,t+1}}{\pi_{t+1}}]$ *and* $1+i_t$ *is the gross nominal rate on a one-period bond,* π_t *the inflation rate and* $f_M = \frac{\partial f}{\partial(M/p)}$. *These two standard conditions are somewhat special in this model. In fact, individual labor supply is downward sloped because it is shifted by changes in the economy-wide labor supply. Decentralizing in a competitive equilibrium and log linearizing the labor market condition we have* $\varphi_c \ln c_t + \varphi_n n_t - (\varphi_N + \varphi_n)\ln N_t = \ln w_t - \ln p_t$. *Since agents are all equal, the aggregate labor supply will be downward sloping function of the real wage and given by* $\varphi_c \ln c_t - \varphi_N \ln N_t = \ln w_t - \ln p_t$. *Hence, a small shift in labor demand increases consumption (which is equal to output in equilibrium) makes real wages fall and employment increase. That is, a demand shock can generate procyclical consumption and employment paths. Also, since money enters the production function, an increase in money could shift labor demand as in the working capital model. However, contrary to that case, labor market effects can be large because of the slope of the aggregate labor supply curve, even when money is relative unimportant as productive factor.*

We will see in exercise 2.35 that there are other more conventional ways to increase output persistence following monetary shocks without using too much price stickiness.

Sticky price models applied to international context produce at least two interesting implications for exchange rate determination and for international risk sharing.

Example 2.11 *(Obstfeld and Rogoff) Consider a structure like the one of example 2.9 where prices are chosen one period in advance, there are two countries , purchasing power parity holds and international financial markets are incomplete, in the sense that only a real bond, denominated in the composite consumption good, is traded. In this economy the domestic nominal interest rate is priced by arbitrage and satisfies* $i + i_{1t} = E_t \frac{p_{1t+1}}{p_{1t}}(i + r_t^B)$,

where r_t^B is the real rate on internationally traded bonds and uncovered interest parity holds, i.e. $1 + i_{1t} = E_t \frac{ner_{t+1}}{ner_t}(1 + i_{2t})$, where $ner_t \equiv \frac{p_{1t}}{p_{2t}}$ and p_{1t} is the consumption based money price index in country 1. Furthermore, the Euler equations for each country imply the international risk sharing condition $E_t[(\frac{c_{1t+1}}{c_{1t}})^{-\varphi} - (\frac{c_{2t+1}}{c_{2t}})^{-\varphi}] = 0$. Hence, while consumption growth needs not be a random walk, the difference in adjusted consumption growth is a martingale difference.

The money demand for each country is $\frac{M_{it+1}}{p_{it}} = \vartheta_m c_{it}(\frac{1+i_{it}}{i_{it}})^{\frac{1}{\varphi_m}}, i = 1, 2$. Hence, using uncovered interest parity and log-linearizing, $\hat{M}_{1t} - \hat{M}_{2t} \propto \frac{1}{\varphi_m}(\hat{c}_{2t} - \hat{c}_{1t}) + \frac{\beta}{(1-\beta)\varphi_m}\widehat{ner}_t$ where $\hat{}$ indicates deviations from the steady state. Hence, differential in money supplies or consumption levels across countries will make the nominal exchange rate jump to a new equilibrium.

Variations or refinements of the price (wage) technology exist in the literature (see Rotemberg (1984) or Dotsey et al. (1999)). Since these refinements are tangential to the scope of this chapter, we invite the interested reader to consult the original sources for details and extensions.

2.2 Approximation methods

As mentioned, finding a solution to the Bellman equation is, in general, complicated. Bellman equation is a functional equation and a fixed point needs to be found in the space of functions. If the regularity conditions for existence and uniqueness are satisfied, this requires iterations which involves the computation of expectations and the maximization of the value function.

We have also seen in example 2.2 and exercise 2.4 that when the utility function is quadratic (logarithmic) and time separable and the constraints are linear in the states and the controls, the form of the value function and of the decision rules is known. In these two situations, if the solution is known to be unique, the method of undetermined coefficients can be used to find the unknown parameters of the solution. Quadratic utility functions are not very appealing, however, as they imply implausible behavior for consumption and asset returns. Log-utility functions are easy to manipulate but they are also restrictive regarding the attitude of agents towards risk. Based on a large body of empirical research, the macroeconomic literature specifies a general power specification for preferences. With this choice one has either to iterate on the Bellman equation or resort to approximations to find a solution.

We have also mentioned that solving general nonlinear expectational equations as those emerging from the first order conditions of a stochastic Lagrangian multiplier problem is complicated. Therefore, also in this case, approximations need to be employed.

This section considers a few approximation methods currently used in the literature. The first approximates the objective function quadratically around the steady state. In the second, the approximation is calculated forcing the states and the exogenous variables to take only a finite number of possible values. This method can be applied to both the value

function and to the first order conditions. The other two approaches directly approximate the optimal conditions of the problem. In one case a log-linear (or a second order) approximation around the steady state is calculated. In the other, the expectational equations are approximated by nonlinear functions and a solution is obtained by finding the parameters of these nonlinear functions.

2.2.1 Quadratic approximations

Quadratic approximations are easy to compute but work under two restrictive conditions. The first is that there exists a point - typically, the steady state - around which the approximation can be taken. Although this requirement may appear innocuous, it should be noted that some models do not possess a steady or a stationary state and in others the steady state may be multiple. The second is that local dynamics are well approximated by linear difference equations. Consequently, such approximations are inappropriate when problems involving large perturbations away from the approximation point (e.g. policy shifts), nonlinear dynamic paths or transitional issues are considered. Moreover, they are incorrect for problems with inequality constraints (e.g. borrowing or irreversibility constraints), since the non-stochastic steady state ignores them.

Quadratic approximations of the objective function are used in situations where the social planner decisions generate competitive equilibrium allocations. When this is not the case the method requires some adaptation to take into account the fact that aggregate variables are distinct from individual ones (see e.g. Hansen and Sargent (1998) or Cooley (1995, chapter 2)) but the same principle works in both cases.

Quadratic approximations can be applied to both value function and Lagrangian multiplier problems. We will discuss applications to the former type problems only since the extension to the other type of problems is straightforward. Let Bellman equation be:

$$\mathbb{V}(y_2, y_3) = \max_{\{y_1\}} u(y_1, y_2, y_3) + \beta E\mathbb{V}(y_2^+, y_3^+ | y_2, y_3) \tag{2.27}$$

where y_2 is a $m_2 \times 1$ vector of the states, y_3 is a $m_3 \times 1$ vector of exogenous variables, y_1 is a $m_1 \times 1$ vector of the controls. Suppose that the constrains are $y_2^+ = h(y_3, y_1, y_2)$ and the law of motion of the exogenous variables is $y_3^+ = \rho_3 y_3 + \epsilon^+$, where h is continuous and ϵ a vector of martingale difference disturbances. Using the constraints into (2.27) we have

$$\mathbb{V}(y_2, y_3) = \max_{\{y_2^+\}} u(y_2, y_3, y_2^+) + \beta E\mathbb{V}(y_2^+, y_3^+ | y_2, y_3) \tag{2.28}$$

Let $\bar{u}(y_2, y_3, y_2^+)$ be the quadratic approximation of $u(y_2, y_3, y_2^+)$ around $(\bar{y}_2, \bar{y}_3, \bar{y}_2)$. If $\mathbb{V}^0$ is quadratic, then (2.28) maps quadratic functions into quadratic functions and the limit value of $V(y_2, y_3)$ will also be quadratic. Hence, under some regularity conditions, the solution to the functional equation is quadratic and the decision rules for y_2^+ linear. When the solution to (2.28) is known to be unique, an approximation to it can be found either iterating on (2.28) starting from a quadratic $\mathbb{V}^0$ or by guessing that $\mathbb{V}(y_2, y_3) = \mathbb{V}_0 + \mathbb{V}_1[y_2, y_3] + [y_2, y_3]\mathbb{V}_2[y_2, y_3]'$, and finding $\mathbb{V}_0, \mathbb{V}_1, \mathbb{V}_2$.

It is important to stress that certainty equivalence is required when computing the solution to a quadratic approximation. This principle allows us to eliminate the expectation operator from (2.28) and reinsert it in front of all future unknown variables once a solution is found. This operation is possible because the covariance matrix of the shocks does not enter the decision rule. That is, certainty equivalence implies that we can set the covariance matrix of the shocks to zero and replace random variables with their unconditional mean.

Exercise 2.20 *Consider the basic RBC model with no habit persistence in consumption and utility given by* $u(c_t, c_{t-1}, N_t) = \frac{c_t^{1-\varphi}}{1-\varphi}$*, no government sector and no taxes and consider the recursive formulation provided by the Bellman equation.*
i) Compute the steady states and a quadratic approximation to the utility function.
iii) Compute the value function assuming that the initial V^0 *is quadratic and calculate the optimal decision rule for capital, labor and consumption (you are supposed to do this by hand but if it becomes overwhelming because of algebra, you can adapt one of the computer programs which comes with this book to undertake the iterations).*

While exercise 2.20 takes a brute force approach to iterations, one should remember that approximate quadratic value function problems fit into the class of optimal linear regulator problems. Therefore, an approximate solution to the functional equation (2.28) can also be found using methods developed in the control literature. One example of an optimal linear regulator problem was encountered in exercise 2.6. The general setup is the following: we want to maximize $E_t \sum_t \beta^t [y_{2t}, y_{3t}]' \mathcal{Q}_2 [y_{2t}, y_{3t}] + y_{1t}' \mathcal{Q}_1 y_{1t} + 2[y_{2t}, y_{3t}]' \mathcal{Q}_3' y_1]$ with respect to y_{1t}, y_{20} given, subject to $y_{2t+1} = \mathcal{Q}_4' y_{2t} + \mathcal{Q}_5' y_{1t} + \mathcal{Q}_6' y_{3t+1}$. Bellman equation is

$$\mathbb{V}(y_2, y_3) = \max_{\{y_1\}} ([y_2, y_3]' \mathcal{Q}_2 [y_2, y_3] + y_1' \mathcal{Q}_1 y_1 + 2[y_2, y_3]' \mathcal{Q}_3' y_1]) + \beta E \mathbb{V}(y_2^+, y_3^+ | y_2, y_3) \quad (2.29)$$

Hansen and Sargent (1998) show that, starting from arbitrary initial conditions, iterations on (2.29) yield at the j-th step, the quadratic value function $\mathbb{V}^j = y_2' \mathbb{V}_2^j y_2 + \mathbb{V}_0^j$ where

$$\mathbb{V}_2^{j+1} = \mathcal{Q}_2 + \beta \mathcal{Q}_4 \mathbb{V}_2^j \mathcal{Q}_4' - (\beta \mathcal{Q}_4 \mathbb{V}_2^j \mathcal{Q}_5' + \mathcal{Q}_3')(\mathcal{Q}_1 + \beta \mathcal{Q}_5 \mathbb{V}_2^j \mathcal{Q}_5')^{-1} (\beta \mathcal{Q}_5 \mathbb{V}_2^j \mathcal{Q}_4' + \mathcal{Q}_3) \quad (2.30)$$

and $\mathbb{V}_0^{j+1} = \beta \mathbb{V}_0^j + \beta tr(\mathbb{V}_2^j \mathcal{Q}_6' \mathcal{Q}_6)$. (2.30) is the so-called matrix Riccati equation which depends on the parameters of the model (i.e. the matrices $\mathcal{Q}_i$), but it does not involve $\mathbb{V}_0^j$. (2.30) can be used to find the limit value $\mathbb{V}_2$ which, in turns, allows us to compute the limit of $\mathbb{V}_0$ and of the value function. The decision rule which attains the maximum at each iteration j is $y_{1t}^j = -(\mathcal{Q}_1 + \beta \mathcal{Q}_5 \mathbb{V}_2^j \mathcal{Q}_5')^{-1} (\beta \mathcal{Q}_5 \mathbb{V}_2^j \mathcal{Q}_4' + \mathcal{Q}_3) y_{2t}^j$ and can be calculated given $\mathbb{V}_2^j$ and the parameters of the model.

While it is common to iterate on (2.30) to find the limits of $\mathbb{V}_0^j, \mathbb{V}_2^j$, the reader should be aware that algorithms which can produce this limit in one step are available (see e.g. Hansen, Sargent and McGrattan (1996)).

Exercise 2.21 *Consider the two country model analyzed in example 2.6.*
i) Take a quadratic approximation to the objective function of the social planner around the steady state and map the problem into a linear regulator problem.
ii) Use the matrix Riccati equation to find a solution to the maximization problem.

The alternative to brute force or Riccati iterations is the method of undetermined coefficients. Although the approach is easy conceptually, it may be mechanically cumbersome, even for small problems. If we knew the functional form of the value function (and/or of the decision rule) we could posit a specific parametric representation and use the first order conditions to solve for the unknown parameters, as we have done in exercise 2.4. We highlight few steps of the approach in the next example and let the reader fill in the details.

Example 2.12 *Suppose the representative agent chooses sequences for* $(c_t, \frac{M_{t+1}}{p_t})$ *to maximize* $E_0 \sum_t \beta^t (c_t^\vartheta + \frac{M_{t+1}}{p_t}^{1-\vartheta})$*, where* c_t *is consumption,* $M_{t+1}^\dagger = \frac{M_{t+1}}{p_t}$ *are real balances, and let* π_t *be the inflation rate. The budget constraint is* $c_t + \frac{M_{t+1}}{p_t} = (1-T^y)w_t + \frac{M_t}{p_t}$*, where* T^y *is an income tax. We assume that* w_t *and* M_t *are exogenous and stochastic. The government budget constraint is* $G_t = T^y w_t + \frac{M_{t+1}-M_t}{p_t}$ *which, together with the consumer budget constraint, implies* $c_t + G_t = w_t$*. Substituting the constraints in the utility function we have* $E_0 \sum_t \beta^t [(1-T^y)w_t + \frac{M_t^\dagger}{\pi_t} + M_{t+1}^\dagger)^\vartheta + (M_{t+1}^\dagger)^{1-\vartheta}$*. The states of the problem are* $y_{2t} = (M_t^\dagger, \pi_t)$ *and the shocks are* $y_{3t} = (w_t, M_t^g)$*. Bellman equation is* $\mathbb{V}(y_2, y_3) = max_{\{c, M^\dagger\}}[u(c, M^\dagger) + \beta E \mathbb{V}(y_2^+, y_3^+ | y_2, y_3)]$*. Let* $(c^{ss}, M^{\dagger ss}, w^{ss}, \pi^{ss})$ *be the steady state value of consumption, real balances, income and inflation. For* $\pi^{ss} = 1, w^{ss} = 1$*, consumption and real balances are given by* $c^{ss} = (1-T^y)$ *and* $(M^\dagger)^{ss} = [\frac{(1-\beta)\vartheta}{1-\vartheta}((1-T^y))^{\vartheta-1}]^{-\frac{1}{\vartheta}}$*. A quadratic approximation to the utility function is* $\mathfrak{B}_0 + \mathfrak{B}_1 x_t + x_t' \mathfrak{B}_2 x_t$ *where* $x_t = (w_t, M_t^\dagger, \pi_t, M_{t+1}^\dagger)$*,* $\mathfrak{B}_0 = (c^{ss})^\vartheta + ((M^\dagger)^{ss})^{1-\vartheta}$*,* $\mathfrak{B}_1 = [\vartheta (c^{ss})^{\vartheta-1}(1-T^y); \frac{\vartheta (c^{ss})^{\vartheta-1}}{\pi^{ss}}; \vartheta (c^{ss})^{\vartheta-1}(-\frac{(M^\dagger)^{ss}}{(\pi^{ss})^2}); -\vartheta (c^{ss})^{\vartheta-1} + (1-\vartheta)((M^\dagger)^{ss})^{-\vartheta}]$ *and the matrix* $\mathfrak{B}_2$ *is*

$$\begin{bmatrix} \kappa(1-T^y)^2 & \frac{\kappa(1-T^y)}{\pi^{ss}} & \kappa((1-T^y)(-\frac{M^{\dagger ss}}{(\pi^{ss})^2}) & -\kappa(1-T^y) \\ \frac{\kappa(1-T^y)}{\pi^{ss}} & \frac{\kappa}{(\pi^{ss})^2} & (-\frac{(M^\dagger)^{ss}}{(\pi^{ss})^2})[\frac{\kappa}{\pi^{ss}} + \vartheta (c^{ss})^{-1})] & -\frac{\kappa}{\pi^{ss}} \\ \kappa(1-T^y)(-\frac{(M^\dagger)^{ss}}{(\pi^{ss})^2}) & (-\frac{(M^\dagger)^{ss}}{(\pi^{ss})^2})[\frac{\kappa}{\pi^{ss}} + \frac{\vartheta}{c^{ss}}] & (-\frac{(M^\dagger)^{ss}}{(\pi^{ss})^2})(\kappa + \frac{\vartheta}{c^{ss}}(-\frac{2}{\pi^{ss}}) & -\frac{(M^\dagger)^{ss}}{(\pi^{ss})^2})(-\kappa) \\ -\kappa(1-T^y) & -\frac{\kappa}{\pi^{ss}} & -\kappa(-\frac{(M^\dagger)^{ss}}{(\pi^{ss})^2})^2 & \kappa + \kappa\frac{((M^\dagger)^{ss})^{\vartheta-1}}{(c^{ss})^{2-\vartheta}} \end{bmatrix}$$

where $\kappa = \vartheta(\vartheta-1)(c^{ss})^{\vartheta-2}$*. One could guess then a quadratic form for the value function and solve for the unknown coefficients. Alternatively, if the decisions rules is everything that is needed, one could directly guess a linear policy function (in deviation from steady states) of the form* $M_{t+1}^\dagger = \mathcal{Q}_0 + \mathcal{Q}_1 M_t^\dagger + \mathcal{Q}_2 w_t + \mathcal{Q}_3 \pi_t + \mathcal{Q}_4 M_t^g$ *and solve for* $\mathcal{Q}_i$ *using the linear version of the first order conditions.*

Exercise 2.22 *Find the approximate first order conditions of the problem of example 2.12. Show the form of* $\mathcal{Q}_j, j = 0, 1, 2, 3$ *(Hint: use the certainty equivalence principle).*

When the number of states is large, analytic calculation of first order and second order derivatives of the return function may take quite some time. As an alternative numerical derivatives, which are much faster to calculate and only require the solution of the model at a pivotal point, could be used. Hence, in example 2.12, one could use e.g. $\frac{\partial u}{\partial c} = \frac{[(1-T^y)w^{ss}+\iota]^\vartheta - [(1-T^y)w^{ss}-\iota]^\vartheta}{2\iota}$, for ι small.

Exercise 2.23 *(Ramsey) Suppose households maximize* $E_0 \sum_t \beta^t (\frac{\upsilon_t c_t^{1-\varphi_c}}{1-\varphi_c} + \frac{N_t^{1-\varphi_n}}{1-\varphi_n})$*, where* υ_t *is a preference shock and* φ_c, φ_n *are parameters. Suppose the resource constraint is* $c_t + G_t = GDP_t = \zeta_t N_t^{\eta}$ *and that the consumer budget constraint is* $E_0 \sum_t \beta^t p_t^0 [(1 - T_t^y) GDP_t + s_{0t}^b - c_t] = 0$*, where* s_{0t}^b *is a stream of coupon payments promised by the government at 0 and* p_t^0 *is an Arrow Debreu price at time zero. The government budget constraint is* $E_0 \sum_t \beta^t p_t^0 [(G_t + s_{0t}^b) - T_t^y GDP_t] = 0$*. Given a process for* G_t *and the present value* $E_0 \sum_t \beta^t p_t^0 s_{0t}^b$*, a feasible tax process must satisfy the government budget constraint. Assume that* $(\upsilon_t, \zeta_t, s_{0t}^b, G_t)$ *are random variables with AR(1) representation. Agents choose sequences for consumption and hours and the government selects the tax process preferred by the representative household. The government commits at time 0 to follow the optimal tax system, once and for all.*
i) Take a quadratic approximation to the problem, calculate the first order conditions of the household problem and show how to calculate the Arrow-Debreu price p_t^0*.*
ii) Show the allocations for c_t, N_t *and the optimal tax policy* T_t^y*. Is it true that the optimal tax rate implies tax smoothing (random walk taxes), regardless of the process for* G_t*?*

Example 2.13 *Consider the setup of exercise 2.8 where the utility function is* $u(c_t, G_t, N_t) = \ln(c_t + \vartheta_G G_t) + \vartheta_N (1 - N_t)$ *and where* G_t *is an AR(1) process with persistence* ρ_G *and variance* σ_G^2 *and it is financed with lump sum taxes. The resource constraint is* $c_t + K_{t+1} + G_t = K_t^{1-\eta} N_t^{\eta} \zeta_t + (1 - \delta) K_t$*, where* $\ln \zeta_t$ *is an AR(1) disturbance with persistence* ρ_ζ *and variance* σ_ζ^2*. Setting* $\vartheta_G = 0.7$*,* $\eta = 0.64, \delta = 0.025, \beta = 0.99, \vartheta_N = 2.8$*, we have that* $(K/GDP)^{ss} = 10.25$*,* $(c/GDP)^{ss} = 0.745, (inv/GDP)^{ss} = 0.225, (G/GDP)^{ss} = 0.03$ *and* $N^{ss} = 0.235$*. Approximating quadratically the utility function and linearly the constraint we can use the matrix Riccati equation to find a solution. Convergence was achieved at iteration 243 and the increment in the value function at the last iteration was 9.41e-06. The value function is proportional to* $[y_2, y_3] \mathbb{V}_2 [y_2, y_3]'$ *where* $y_2 = K; y_3 = (G, \zeta)$ *and* $\mathbb{V}_2 = \begin{bmatrix} 1.76e-09 & 3.08e-07 & 7.38e-09 \\ -1.54e-08 & -0.081 & -9.38e-08 \\ -2.14e-06 & -3.75e-04 & -8.98e-06 \end{bmatrix}$*. The decision rule for* $y_1 = (c, N)'$ *is* $y_{1t} = \begin{bmatrix} -9.06e-10 & -0.70 & -2.87e-09 \\ -9.32e-10 & -1.56e-07 & -2.95e-09 \end{bmatrix} y_{2t}$.

2.2.2 Discretization

As an alternative to quadratic approximations, one could solve the value function problem by discretizing the state space and the space over which the exogenous processes take values. This is the method popularized, for example, by Merha and Prescott (1985). The idea is that the states are forced to lie in the set $Y_2 = \{y_{21}, \ldots y_{2n_1}\}$ and the exogenous processes in the set $Y_3 = \{y_{31}, \ldots y_{3n_2}\}$. Then the space of possible (y_{2t}, y_{3t}) combination has $n_1 \times n_2$ points. For simplicity, assume that the process for the exogenous variables is first order Markov with transition $P(y_{3t+1} = y_{3j'} | y_{3t} = y_{3j}) = p_{j'j}$. The value function associated with each pair of states and exogenous processes is $\mathbb{V}(y_{2i}, y_{3j})$, which is of dimension $n_1 \times n_2$. Because of the Markov structure of the shocks, and the assumptions made, we have transformed an

infinite dimensional problem into the problem of mapping of $n_1 \times n_2$ matrices into $n_1 \times n_2$ matrices. Therefore, iterations on the Bellman equation are easier to compute. The value function can be written as $(\mathcal{T}\mathbb{V})(y_{2i}, y_{3j}) = \max_n u(y_1, y_{2i}, y_{3j}) + \beta \sum_{l=1}^{n_2} \mathbb{V}_{n,l} p_{l,j}$, where y_{1n} is such that $h(y_{1n}, y_{2i}, y_{3j}) = y_{2n}$, $n = 1, \ldots n_1 \times n_2$. An illustration of the approach is given in the next example.

Example 2.14 *Consider a RBC model where a random stream of government expenditure is financed by distorting income taxes, labor supply is inelastic and production uses only capital. The social planner chooses* $\{c_t, K_{t+1}\}$ *to maximize* $E_0 \sum_t \beta^t \frac{c_t^{1-\varphi}}{1-\varphi}$, *given* G_t *and* K_t, *subject to* $c_t + K_{t+1} - (1-\delta)K_t + G_t = (1-T^y)K_t^{1-\eta}$, *where* G_t *is an AR(1) with persistence* ρ_g, *variance* σ_g^2 *and* $(\varphi, \beta, T^y, \eta, \delta)$ *are parameters. Bellman equation is* $\mathbb{V}(K,G) = \max_{\{K^+\}} \frac{[(1-T^y)K^{1-\eta} + (1-\delta)K - G - K^+]^{1-\varphi}}{1-\varphi} + \beta E(\mathbb{V}(K^+, G^+|K, G))$. *Given* K_0, *define* $\mathcal{T}$ *by* $(\mathcal{T}\mathbb{V})(K,G) = \max_{\{K^+\}} \frac{[(1-T^y)K^{1-\eta} + (1-\delta)K - G - K^+]^{1-\varphi}}{1-\varphi} + \beta E(\mathbb{V}(K^+, G^+|K, G)$. *Suppose that the capital stock and government expenditure can take only two values, and let the transition for* G_t *be* $p_{j'j}$. *Then the discretization algorithm works as follows:*

Algorithm 2.2

1) *Choose values for* $(\delta, \eta, \varphi, T^y, \beta)$ *and specify the elements of* $p_{j,j'}$.

2) *Choose an initial* 2×2 *matrix* $\mathbb{V}(K,G)$, *e.g.* $\mathbb{V}^0 = 0$.

3) *For* $i, i', j, j' = 1, 2$, *calculate* $(\mathcal{T}\mathbb{V}_{i,i'})(K,G) = \max([\frac{(1-T^y)K_i^{1-\eta} + (1-\delta)K_i - K_i - G_{i'})^{1-\vartheta}}{1-\vartheta}] + \beta [\mathbb{V}_{i,i'} p_{i,i'} + \mathbb{V}_{i,j'} p_{i,j'})]$; $[\frac{(1-T^y)K_i^{1-\eta} + (1-\delta)K_i - K_j - G_{i'})^{1-\vartheta}}{1-\vartheta}] + \beta[\mathbb{V}_{j,i'} p_{j,i'} + \mathbb{V}_{j,j'} p_{j,j'})])$.

4) *Iterate on 3) until e.g.* $max_{i,i'} |\ \mathcal{T}^l \mathbb{V}_{i,i'} - \mathcal{T}^{l-1} \mathbb{V}_{i,i'}| \leq \iota$, ι *small,* $l = 2, 3, \ldots$.

Suppose $T^y = 0.1$, $\delta = 0.1; \beta = 0.9, \varphi = 2, \eta = 0.66$, *choose* $G_1 = 1.1; G_2 = 0.9$, $K_1 = 5.3$, $K_2 = 6.4$, $p_{11} = 0.8$; $p_{22} = 0.7$, $\mathbb{V}^0 = 0$.
Then $(\mathcal{T}\mathbb{V})_{11} = max_{1,2}\{(\frac{(1-T^y)K_1^{1-\eta} + (1-\delta)K_1 - K_1 - G_1)^{1-\varphi}}{1-\varphi}); (\frac{(1-T^y)K_1^{1-\eta} + (1-\delta)K_1 - K_2 - G_1)^{1-\vartheta}}{1-\vartheta})\}$ $= max_{1,2}(14.38, 0.85) = 14.38$. *Repeating for the other entries,* $\mathcal{T}\mathbb{V} = \begin{bmatrix} 14.38 & 1.03 \\ 12.60 & -0.81 \end{bmatrix}$; $\mathcal{T}^2\mathbb{V} = \begin{bmatrix} 24.92 & 3.91 \\ 21.53 & 1.10 \end{bmatrix}$, *and* $\lim_{l\to\infty} \mathcal{T}^l\mathbb{V} = \begin{bmatrix} 71.63 & 31.54 \\ 56.27 & 1.10 \end{bmatrix}$. *Implicitly the solution defines the decision rule; for example, from* $(\mathcal{T}\mathbb{V})_{11}$ *we have that* $K_t = K_1$.

Clearly, the quality of the approximation depends on the finess of the grid. Therefore, it is a good idea to start from course grids and after convergence is achieved check whether a finer grid in the space of states or in the space of shocks produces different results.

The discretization approach is well suited for problems of modest dimension (i.e. when the number of state variables and of exogenous processes is small) since constructing a grid which systematically and effectively covers high dimensional spaces is difficult. For example,

when we have one state, two shocks and 100 grid points, 1,000,000 evaluations are required in each step. Nevertheless, even with this large number of evaluations, it is easy to leave large portions of the space unexplored.

Exercise 2.24 *(Search) Suppose an agent has the choice of accepting or rejecting a wage offer. If she has worked at $t-1$, the offer is $w_t = b_0 + b_1 w_{t-1} + e_t$, where e_t is a shock; if she was searching at $t-1$, the offer is drawn from some stationary distribution. Having observed w_t agents decide whether to work or not (i.e. whether $N_t = 0$ or $N_t = 1$). Agents can't save so $c_t = w_t$ if $N_t = 1$ and $c_t = \bar{c}$ if $N_t = 0$, where $\bar{c}$ measures unemployment compensations. Agents maximize discounted utility where $u(c) = \frac{c_t^{1-\varphi}}{1-\varphi}$ and φ is a parameter.*
i) Write down the maximization problem and the first order conditions.
ii) Define states and controls and write down the Bellman equation. Suppose $e_t = 0, b_0 = 0, b_1 = 1$; $\beta = 0.96$ and $w_t \sim \mathbb{U}(0,1)$. Calculate the optimal value function and describe the decision rules.
iii) Assume that the agent has now also the option of retiring so that $x_t = 0$ or $x_t = 1$. Suppose $x_t = x_{t-1}$ if $x_{t-1} = 0$ and that $c_t = w_t$ if $N_t = 1$, $x_t = 1$; $c_t = \bar{c}$ if $N_t = 0$, $x_t = 1$ and $c_t = \bar{\bar{c}}$ if $N_t = 0, x_t = 0$, where $\bar{\bar{c}}$ is the retirement pay. Write down Bellman equation and calculate the optimal decision rules.
iv) Suppose that the agent has now the option to migrate. For each location $i = 1, 2$ the wage is $w_t^i = b_0 + b_1 w_{t-1}^l + e_t^i$ if the agent has worked at $t-1$ in location i, and by $w_t^i \sim \mathbb{U}(0, i)$ otherwise. Consumption is $c_t = w_t$ if $i_t = i_{t-1}$ and $c_t = \bar{c} - \varrho$ if $i_t \neq i_{t-1}$, where $\varrho = 0.1$ is a migration cost. Write down the Bellman equation and calculate the optimal decision rules.

Exercise 2.25 *(Lucas tree model) Consider an economy where infinitely lived agents have a random stream of perishable endowments sd_t and decide how much to consume and save, where savings can take the form of either stocks or bonds and let $u(c_t, c_{t-1}, N_t) = \ln c_t$*
i) Write down the maximization problem and the first order conditions. Write down the Bellman equation specifying the states and the controls.
ii) Assume that the endowment process can take only two values $sd_1 = 6, sd_2 = 1$ with transition $\begin{bmatrix} 0.7 & 0.3 \\ 0.2 & 0.8 \end{bmatrix}$. Find the 2×1 vector of value functions, one for each state.
iii) Find the policy function for consumption, stock and bond holdings and the pricing functions for stocks and bonds.

One can employ a discretization approach also to solve the optimality conditions of the problem. Hence, the methodology is applicable to problems where the value function may not exist.

Example 2.15 *For general preferences, the Euler equation of exercise 2.25 is*

$$p_t^s(sd_t) U_{c,t} = \beta E[U_{c,t+1}(p_{t+1}^s(sd_{t+1}) + sd_{t+1})] \tag{2.31}$$

where we have made explicit the dependence of p_t^s on sd_t. If we assume that $sd_t = [sd_h, sd_L]$, use the equilibrium condition $c_t = sd_t$ and let $U_i^1 \equiv p^s(sd_i) U_{sd_i}$; $U_i^2 = \beta \sum_{i'=1}^{2} sd_i U_{sd_i} p_{ii'}$,

(2.31) can be written as $U_i^1 = U_i^2 + \beta \sum_{i'} p_{ii'} U_{i'}^1$ *or* $U^1 = (1 - \beta P)^{-1} U^2$*, where* P *is the matrix with typical element* $\{p_{ij}\}$*. Therefore, given a functional form for the utility, share prices satisfy* $p^s(sd_i) = \sum_{i'} (I + \beta P + \beta^2 P^2 + \ldots)_{ii'} \frac{U_{i'}^2}{U_{sd_i}}$*, where the sum is over the* (i, i') *elements of the matrix.*

Exercise 2.26 *Consider the interteporal condition (2.13), the intratemporal condition (2.14) of a standard RBC economy. Assume* $T^y = 0$ *and that* (K_t, ζ_t) *can take two values. Describe how to find the optimal consumption/leisure choice when* $U(c_t, c_{t-1}, N_t) = \ln c_t + \vartheta_N (1 - N_t)$*.*

2.2.3 Log linear Approximations

Log linearizations have been extensively used in recent years following the work of Blanchard and Khan (1980), King, Plosser and Rebelo (1987) and Campbell (1994). Uhlig (1999) has systematized the methodology and provided software useful to solve a variety of problems. King and Watson (1998) and Klein (2000) provided algorithms for singular systems and Sims (2001) for problems where the distinction between states and controls is unclear.

Log linear approximations are similar, in spirit, to quadratic approximations and the solutions are computed using similar methodologies. The former may work better when the problem displays some mild non-linearities. The major difference between the two approaches is that quadratic approximations are typically performed on the objective function while log-linear approximations are calculated using the optimality conditions of the problem. Therefore, the latters are useful in situations where, because of distortions, the competitive equilibrium is suboptimal.

The basic principles of log linearization are simple. We need a point around which the log-linearization takes place. This could be the steady state or, in models with frictions, the frictionless solution. Let $y = (y_1, y_2, y_3)$. The optimality conditions of the problem can be divided into two blocks, the first containing expectational equations and the second non-expectational equations:

$$1 = E_t[h(y_{t+1}, y_t)] \tag{2.32}$$

$$1 = f(y_t, y_{t-1}) \tag{2.33}$$

where $f(0,0) = 1;\ h(0,0) = 1$. Taking first order Taylor expansions around $(\bar{y}, \bar{y}) = (0, 0)$

$$0 \approx E_t[h_{t+1} y_{t+1} + h_t y_t] \tag{2.34}$$

$$0 = f_t y_t + f_{t-1} y_{t-1} \tag{2.35}$$

where $f_j = \frac{\partial \ln f}{\partial y_j'}$ and $h_j = \frac{\partial \ln h}{\partial y_j'}$. (2.34) and (2.35) form a system of linear expectational equations.

Although log linearization only requires calculations of the first derivatives of the functions f and h, Uhlig (1999) suggests a set of approximations to calculate (2.34)-(2.35) directly without differentiation. The tricks involve replacing any y_t with $\bar{y} e^{\tilde{y}_t}$, where $\tilde{y}_t$ is small and using the following three rules (here a_0 is a constant and b_{1t}, b_{2t} small numbers).

- $e^{b_{1t}+a_0 b_{2t}} \approx 1 + b_{1t} + a_0 b_{2t}$.
- $b_{1t} b_{2t} \approx 0$.
- $E_t[a_0 e^{b_{1t+1}}] \propto E_t[a_0 b_{1t+1}]$.

Example 2.16 *To illustrate how to use these rules consider the resource constraint* $C_t + G_t + Inv_t = GDP_t$. *Set* $\bar{c}e^{c_t} + \bar{g}e^{g_t} + \overline{inv}e^{inv_t} = \overline{GDP}e^{gdp_t}$ *and use rule i) to get* $\bar{c}(1+c_t) + \bar{g}(1+g_t) + \delta\bar{k}(1+inv_t) - \overline{GDP}(1+gdp_t) = 0$. *Then using* $\bar{c} + \bar{g} + \overline{inv} = \overline{GDP}$ *we get* $\bar{c}c_t + \bar{g}g_t + \overline{inv}inv_t - \overline{GDP}gdp_t = 0$ *or* $\frac{\bar{c}}{\overline{GDP}}c_t + \frac{\bar{g}}{\overline{GDP}}g_t + \frac{\overline{inv}}{\overline{GDP}}inv_t - gdp_t = 0$.

Exercise 2.27 *Suppose* y_t *and* y_{t+1} *are conditionally jointly log-normal homoschedastic processes. Replace (2.32) with* $0 = log\{E_t[exp(\bar{h}(y_{t+1}, y_t))]\}$, *where* $\bar{h} = log(h)$. *Using* $\log h(0,0) \approx 0.5 var_t[\bar{h}_{t+1}y_{t+1} + \bar{h}_t y_t]$, *show that the log linear approximation is* $0 \approx E_t[\bar{h}_{t+1}y_{t+1}$
$+ \bar{h}_t y_t)]$. *What is the difference between this approximation and the one in (2.34)?*

Exercise 2.28 *Suppose that the private production is* $GDP_t = K_t^{1-\eta} N_t^{\eta} \zeta_t (\frac{K_t}{Pop_t})^{\aleph_1/(1-\eta)} (\frac{N_t}{Pop_t})^{\aleph_2/\eta}$ *where* $(\frac{K_t}{Pop_t})$ *and* $(\frac{N_t}{Pop_t})$ *are the average endowment of capital and hours in the economy. Suppose the utility function is* $E_t \sum_t \beta^t [\ln(\frac{c_t}{Pop_t}) - \frac{1}{1-\varphi_N}(\frac{N_t}{Pop_t})^{1-\varphi_N}]$. *Assume that* $(\ln \zeta_t, \ln Pop_t)$ *are AR(1) processes with persistence equal to* ρ_ζ *and* 1.
i) Show that the optimality conditions of the problem are

$$\frac{c_t}{Pop_t}(\frac{N_t}{Pop_t})^{-\varphi_N} = \eta \frac{GDP_t}{Pop_t} \tag{2.36}$$

$$\frac{Pop_t}{c_t} = E_t \beta \frac{Pop_{t+1}}{c_{t+1}}[(1-\delta) + (1-\eta)\frac{GDP_{t+1}}{K_{t+1}}] \tag{2.37}$$

ii) Find expressions for the log-linearized production function, the labor market equilibrium, the Euler equation and the budget constraint.
iii) Write the log linearized expectational equation in terms of an Euler equation error. Give conditions under which sunspot equilibria may obtain (Hint: find conditions under which there are more stable roots than state variables).

There are several economic models which do not fit the setup of (2.32)-(2.33). For example, Rotemberg and Woodford (1997) describe a model where consumption at time t depends on the expectations of variables dated at $t+2$ and on. This model can be accommodated in the setup of (2.32)-(2.33) using dummy variables, as the next example, shows. In general, restructuring of the timing convention of the variables, or enlarging the vector of states, suffices to fit these problems into (2.32)- (2.33).

Example 2.17 *Suppose that (2.32) is* $1 = E_t[h(y_{2t+2}, y_{2t})]$. *We can transform this second order expectational equation into a* 2×1 *vector of first order expectational equations using a dummy variable* y^*_{2t}. *In fact, the above is equivalent to* $1 = E_t[h(y^*_{2t+1}, y_{2t})]$ *and* $0 = E^*_t(y_{2t+1}, y^*_{2t})$ *as long as the vector* $[y_{2t}, y^*_{2t}]$ *is used as state variables for the problem.*

Exercise 2.29 *Consider a model with optimizers and rule of thumb consumers like the one of example 2.7 and assume that optimizing agents display habit in consumption. In particular, assume that their utility function is* $(c_t - \gamma c_{t-1})^{\vartheta}(1-N_t)^{1-\vartheta}$. *Derive the first order conditions of the model and map them into (2.32)-(2.33).*

Example 2.18 *Log-linearizing around the steady state the equilibrium conditions of the model of exercise 2.14, and assuming an unexpected change in the productivity of farmers technology (represented by* Δ*) lasting one period we have:* $(1+\frac{1}{\varrho})\widehat{La}_t = \Delta + \frac{r}{r-1}\hat{p}_t^L$ *for* $\tau = 0$ *and* $(1+\frac{1}{\varrho})\widehat{La}_{t+\tau} = \widehat{La}_{t+\tau-1}$ *for* $\tau \geq 1$ *where* ϱ *is the elasticity of the supply of land with respect to the user costs in the steady state and* $\hat{p}_t^L = \frac{r-1}{r\varrho}\frac{1}{1-\frac{\varrho}{r(1+\varrho)}}\widehat{La}_t$, *where* $\hat{}$ *indicates percentage deviations from the steady state. Solving these two expressions we have* $\hat{p}_t^L = \frac{\Delta}{\varrho}$ *and* $\widehat{La}_t = \frac{1}{1+\frac{1}{\varrho}}(1+\frac{r}{(r-1)\varrho})\Delta$. *Three interesting conclusions follows. First, if* $\varrho = 0$, *temporary shocks have permanent effects on farmers land and on its price. Second, since* $\frac{1}{1+\frac{1}{\varrho}}(1+\frac{r}{(r-1)\varrho}) > 1$, *the effect on land ownership is larger than the shock. Finally, in the static case* $(\hat{p}_t^L)^* = \frac{r-1}{r\varrho}\Delta < \hat{p}_t^L$ *and* $(\widehat{La}_t)^* = \Delta < \widehat{La}_t$. *This is because* Δ *affects the net worth of farmers: a positive* Δ *reduces the value of the obligations and implies a larger use of capital by the farmers, therefore magnifying the effect of the shock on land ownership.*

Exercise 2.30 *Show that the log-linearized first order conditions of the sticky price model of example 2.9 when* $K_t = 1, \forall t$ *are*

$$\begin{aligned}
0 &= w_t + \frac{N^{ss}}{1-N^{ss}}N_t - c_t \\
(\frac{1}{1+i^{ss}})i_{t+1} &= (1-\vartheta(1-\varphi))(c_{t+1}-c_t) - (1-\vartheta)(1-\varphi)(N_{t+1}-N_t)\frac{N^{ss}}{1-N^{ss}} - \pi_{t+1} \\
\frac{M_{t+1}}{p_t} &= \frac{\vartheta(1-\varphi)-1}{\varphi_m}c_t + \frac{N^{ss}}{1-N^{ss}}\frac{(1-\vartheta)(1-\varphi)}{\varphi_m}N_t - \frac{1}{\varphi_m(1+i^{ss})}i_t \\
\beta E_t\pi_{t+1} &= \pi_t - \frac{(1-\zeta_p)(1-\zeta_p\beta)}{\zeta_p}mc_t
\end{aligned} \tag{2.38}$$

where mc_t *are real marginal costs,* ζ_p *is the probability of not changing the prices,* w_t *is the real wage,* φ *is the risk aversion parameter,* ϑ *is the share of consumption in utility,* ϑ_M *is the exponent on real balances in utility and the superscript* ss *refers to steady state.*

As with quadratic approximations, the solution of the system of equations (2.34)-(2.35) can be obtained in two ways when the solution is known to exist and be unique: using the method of the undetermined coefficients or finding the saddle-point solution (Vaughan's method). The method of undetermined coefficients is analogous to the one described in exercise 2.20. Vaughan's method works with the state space representation of the system. Both methods require the computation of eigenvalues and eigenvectors. For a thorough discussion of the methods and a comparison with second order difference equation methods, the reader should consult e.g. the chapter of Uhlig in Marimon and Scott (1999) or

Klein (2000). Here we briefly describe the building blocks of the procedure and highlight important steps with some examples.

Rather than using (2.34) and (2.35), we employ a slightly more general setup which directly allows for structures like those considered in exercises **??** and 2.29, without any need to enlarge the state space.

Let y_{1t} be of dimension $m_1 \times 1$, y_{2t} of dimension $m_2 \times 1$, and y_{3t} of dimension $m_3 \times 1$ and suppose the log linearized first order conditions, the budget constraint and the law of motion of the exogenous variables be written as:

$$\begin{aligned} 0 &= \mathcal{Q}_1 y_{2t} + \mathcal{Q}_2 y_{2t-1} + \mathcal{Q}_3 y_{1t} + \mathcal{Q}_4 y_{3t} && (2.39)\\ 0 &= E_t[\mathcal{Q}_5 y_{2t+1} + \mathcal{Q}_6 y_{2t} + \mathcal{Q}_7 y_{2t-1} + \mathcal{Q}_8 y_{1t+1} + \mathcal{Q}_9 y_{1t} + \mathcal{Q}_{10} y_{3t+1} + \mathcal{Q}_{11} y_{3t}] && (2.40)\\ 0 &= y_{3t+1} - \rho y_{3t} - \epsilon_t && (2.41) \end{aligned}$$

where $\mathcal{Q}_3$ is a $m_4 \times m_1$ matrix and of rank $m_1 \leq m_4$, and ρ has only stable eigenvalues. Assume that a solution is given by:

$$\begin{aligned} y_{2t} &= \mathcal{A}_{22} y_{2t-1} + \mathcal{A}_{23} y_{3t} && (2.42)\\ y_{1t} &= \mathcal{A}_{12} y_{2t-1} + \mathcal{A}_{13} y_{3t} && (2.43) \end{aligned}$$

Letting $\mathcal{Z}_1 = \mathcal{Q}_8 \mathcal{Q}_3^+ \mathcal{Q}_2 - \mathcal{Q}_6 + \mathcal{Q}_9 \mathcal{Q}_3^+ \mathcal{Q}_1$, Uhlig (1999) shows that:

a) $\mathcal{A}_{22}$ satisfies the (matrix) quadratic equations:

$$\begin{aligned} 0 &= \mathcal{Q}_3^0 \mathcal{Q}_1 \mathcal{A}_{22} + \mathcal{Q}_3^0 \mathcal{Q}_2 && (2.44)\\ 0 &= (\mathcal{Q}_5 - \mathcal{Q}_8 \mathcal{Q}_3^+ \mathcal{Q}_1)\mathcal{A}_{22}^2 - \mathcal{Z}_1 \mathcal{A}_{22} - \mathcal{Q}_9 \mathcal{Q}_3^+ \mathcal{Q}_2 + \mathcal{Q}_7 && (2.45) \end{aligned}$$

The equilibrium is stable if all eigenvalues of $\mathcal{A}_{22}$ are less than one in absolute value.

b) $\mathcal{A}_{12}$ is given by $\mathcal{A}_{12} = -\mathcal{Q}_3^+(\mathcal{Q}_1 \mathcal{A}_{22} + \mathcal{Q}_2)$.

c) Given $\mathcal{Z}_2 = (\mathcal{Q}_5 \mathcal{A}_{22} + \mathcal{Q}_8 \mathcal{A}_{12})$ and $\mathcal{Z}_3 = \mathcal{Q}_{10}\rho + \mathcal{Q}_{11}$, $\mathcal{A}_{13}$ and $\mathcal{A}_{23}$ satisfy:

$$\begin{bmatrix} I_{m_3} \otimes \mathcal{Q}_1 & I_{m_3} \otimes \mathcal{Q}_3 \\ \rho' \otimes \mathcal{Q}_5 + I_{m_3} \otimes (\mathcal{Z}_2 + \mathcal{Q}_6) & \rho' \otimes \mathcal{Q}_8 + I_{m_3} \otimes \mathcal{Q}_9 \end{bmatrix} \begin{bmatrix} vec(\mathcal{A}_{23}) \\ vec(\mathcal{A}_{13}) \end{bmatrix} = - \begin{bmatrix} vec(\mathcal{Q}_4) \\ vec(\mathcal{Z}_3) \end{bmatrix}$$

where $vec(.)$ is columnwise vectorization; $\mathcal{Q}_3^G$ is a pseudo inverse of $\mathcal{Q}_3$ and satisfies $\mathcal{Q}_3^G \mathcal{Q}_3 \mathcal{Q}_3^G = \mathcal{Q}_3^G$ and $\mathcal{Q}_3 \mathcal{Q}_3^G \mathcal{Q}_3 = \mathcal{Q}_3$; $\mathcal{Q}_3^0$ is an $(m_4 - m_1) \times m_4$ matrix whose rows are a basis for the space of $\mathcal{Q}_3'$ and I_{m_3} is the identity matrix of dimension m_3.

Example 2.19 *Consider a RBC model with an intermediate monopolistic competitive sector. Let the profits in firm i be $prf_{it} = mk_t inty_t$, where $mk_t = (p_{it} - mc_{it})$ is the markup. If the utility function is of the form $u(c_t, c_{t-1}, N_t) = \frac{c_t^{1-\varphi}}{1-\varphi} + \vartheta_N(1 - N_t)$, the dynamics depend on the markup only via the steady state. For this model the log linearized conditions are*

$$0 = -(inv^{ss}/GDP^{ss})inv_t - (c^{ss}/GDP^{ss})c_t + gdp_t \qquad (2.46)$$

$$
\begin{aligned}
0 &= (inv^{ss}/K^{ss})inv_t - k_{t+1} + (1-\delta)k_t && (2.47)\\
0 &= (1-\eta)k_t - gdp_t + \eta n_t + \zeta_t && (2.48)\\
0 &= -\varphi c_t + gdp_t - N_t && (2.49)\\
0 &= -k_t + gdp_t - \frac{r^{ss}}{mk^{ss}(1-\eta)(GDP^{ss}/K^{ss})} r_t && (2.50)\\
0 &= E_t[-\varphi c_{t+1} + r_{t+1} + \varphi c_t] && (2.51)\\
\zeta_{t+1} &= \rho_\zeta \zeta_t + \epsilon_{1t+1} && (2.52)
\end{aligned}
$$

where (inv^{ss}/GDP^{ss}) *and* (c^{ss}/GDP^{ss}) *are the steady state investment and consumption to output ratios,* r^{ss} *is the steady state real rate and* mk^{ss} *steady state markup. Letting* $y_{1t} = (c_t, gdp_t, N_t, r_t, inv_t)$, $y_{2t} = k_t$, $y_{3t} = \zeta_t$, *we have* $\mathcal{Q}_5 = \mathcal{Q}_6 = \mathcal{Q}_7 = \mathcal{Q}_{10} = \mathcal{Q}_{11} = [0]$

$$
\mathcal{Q}_2 = \begin{bmatrix} 0 \\ (1-\delta)K^{ss} \\ 1-\eta \\ 0 \\ -D^{ss} \end{bmatrix};\ \mathcal{Q}_3 = \begin{bmatrix} -C^{ss} & GDP^{ss} & 0 & 0 & -inv^{ss} \\ 0 & 0 & 0 & 0 & inv^{ss} \\ 0 & -1 & \eta & 0 & 0 \\ -\varphi & 1 & -1 & 0 & 0 \\ 0 & D^{ss} & 0 & -r^{ss} & 0 \end{bmatrix};\ \mathcal{Q}_1 = \begin{bmatrix} 0 \\ -K^{ss} \\ 0 \\ 0 \\ 0 \end{bmatrix};
$$

$$
\mathcal{Q}_4 = \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \\ 0 \end{bmatrix};\ \mathcal{Q}_8 = [-\varphi, 0, 0, 1, 0];\ \mathcal{Q}_9 = [\varphi, 0, 0, 0, 0], \rho = [\rho_\zeta];\ \textit{where}\ D^{ss} = mk^{ss}(1-\eta)(GDP^{ss}/K^{ss}).
$$

It is important to stress that the method of undetermined coefficients properly works only when the state space is chosen to be of minimal size, that is, no redundant state variables are included. If this is not the case, $\mathcal{A}_{22}$ may have zero eigenvalues and this will produce "bubble" solutions.

Computationally, the major difficulty is to find a solution to the matrix equation (2.45). The toolkit of Uhlig (1999) recasts it into a generalized eigenvalue-eigenvector problem. Klein (2000) and Sims (2001) calculate a solution using the generalized Shur decomposition. When applied to some of the problems described in this chapter, the two approaches yield similar solutions. In general, the Shur (QZ) decomposition is useful when generalized eigenvalues may not be distinct. However, the QZ decomposition is not necessarily unique.

Exercise 2.31 *Suppose the representative agent maximizes* $E_0 \sum_t \beta^t \frac{c_t^{1-\varphi_c}}{1-\varphi_c} + \frac{(\frac{M_{t+1}}{p_t})^{1-\varphi_m}}{1-\varphi_m}$, *where* φ_c *and* φ_m *are parameters, subject to the resource constraint* $c_t + K_{t+1} + \frac{M_{t+1}}{p_t} = \zeta_t K_t^{1-\eta} N_t^\eta + (1-\delta)K_t + \frac{M_t}{p_t}$ *where* $\ln \zeta_t$ *is an AR(1) process with persistence* ρ_ζ *and standard error* σ_ζ. *Let* $M_{t+1}^\dagger = \frac{M_{t+1}}{p_t}$ *be real balances,* π_t *the inflation rate,* r_t *the rental rate of capital and assume* $\ln M_{t+1}^s = \ln M_t^s + \ln M_t^g$ *where* $\ln M_t^g$ *has mean* $\bar{M} \geq 0$ *and standard error* σ_M. *i) Verify that the first order conditions of the problem are*

$$
r_t = (1-\eta)\zeta_t K_t^{-\eta} N_t^\eta + (1-\delta)
$$

$$\begin{aligned} 1 &= E_t[\beta(\frac{c_{t+1}}{c_t})^{-\varphi_c} r_{t+1}] \\ (M^{\dagger}_{t+1})^{-\vartheta_m} c_t^{-\varphi_c} &= 1 + E_t[\beta(\frac{c_{t+1}}{c_t})^{-\varphi_c} \pi_{t+1}] \end{aligned} \tag{2.53}$$

ii) Log linearize (2.53), the resource constraint, and the law of motion of the shocks and cast these equations into the form of equations (2.39)-(2.41).
iii) Guess that a solution for $[K_{t+1}, c_t, r_t, M^{\dagger}_{t+1}]$ *is linear in* $(K_t, M^{\dagger}_t, \zeta_t, M^g_t)$*. Determine the coefficients of the relationship.*

Exercise 2.32 *Suppose agents maximize* $E_0 \sum_{t=0}^{\infty} \beta^t u(c_t, 1 - N_t)$ *subject to* $c_t + \frac{M_{t+1}}{p_t} + K_{t+1} \leq (1-\delta)K_t + (GDP_t - G_t) + \frac{M_t}{p_t} + T_t$, $\frac{M_t}{p_t} \geq c_t$, *where* $GDP_t = \zeta_t K_t^{1-\eta} N_t^{\eta}$ *and assume that the monetary authority sets* $\Delta \ln M^s_{t+1} = \ln M^g_t + a i_t$, *where* a *is a parameter and* i_t *the nominal interest rate. The government budget constraint is* $G_t + \frac{M_{t+1} - M_t}{p_t} = T_t$. *Let* $[\ln G_t, \ln \zeta_t, \ln M^g_t]$ *be a vector of random disturbances.*
i) Assume a binding CIA constraint, $c_t = \frac{M_{t+1}}{p_t}$. *Derive the optimality conditions and the equation determining the nominal interest rate.*
ii) Compute a log-linear approximation around the steady states of the first order conditions and of the budget constraint, of the production function, of the CIA constraint, of the equilibrium pricing equation for nominal bonds and the of government budget constraint.
iii) Show that the system is recursive and can be solved for $(N_t, K_t, \frac{M_{t+1}}{p_t}, i_t)$ *first while* $(GDP_t, c_t, \lambda_t, T_t)$ *can be solved in a second stage as a function of* $(N_t, K_t, \frac{M_{t+1}}{p_t}, i_t)$, *where* λ_t *is the Lagrangian multiplier on the private budget constraint.*
iv) Write down the system of difference equations for $(N_t, K_t, \frac{M_t}{p_t}, i_t)$. *Guess a linear solution (in deviation from steady states) in* K_t *and* $[\ln G_t, \ln \zeta_t, \ln M^g_t]$ *and find the coefficients.*
v) Assume prices are set one period in advance as a function of the states and of past shocks, i.e. $p_t = a_0 + a_1 K_t + a_{21} \ln G_{t-1} + a_{22} \ln \zeta_{t-1} + a_{23} \ln M^g_{t-1}$. *What is the state vector in this case? What is the most likely guess for* $(N_t, K_t, \frac{M_t}{p_t}, i_t)$*? Use the method of undetermined coefficients to find a solution.*

The next example shows the form of the log-linearized solution of a version of the sticky price-sticky wage model described in exercise 2.18.

Example 2.20 *Assume that capital is fixed so that the only variable factor in production is labor and the utility function is* $E_0 \sum_t \beta^t (\frac{c_t^{\vartheta}(1-N_t)^{1-\vartheta})^{1-\varphi}}{1-\varphi} + \frac{\vartheta_m}{1-\varphi_m}(\frac{M_{t+1}}{p_t})^{1-\varphi_m})$. *Set* $N^{ss} = 0.33, \eta = 0.66$, $\pi^{ss} = 1.005$, $\beta = 0.99$, $(\frac{c}{GDP})^{ss} = 0.8$, *where* $(\frac{c}{GDP})^{ss}$ *is the share of consumption in GDP,* N^{ss} *is hours worked and* π^{ss} *is gross inflation in the steady states,* η *is exponent of labor in the production function,* β *is the discount factor. These choices imply, for example, that in steady-state the gross real interest rate is 1.01, output is 0.46, real balances 0.37 and the real (fully flexible) wage 0.88. We select the degree of price and wage rigidity to be the same and set* $\zeta_p = \zeta_w = 0.75$. *Given the quarterly frequency of the model, this choice implies that on average firms (consumers) change their price (wage) every three quarters. Also, we choose the elasticity of money demand* $\vartheta_m = 7$. *In the monetary*

policy rule we set $a_2 = -1.0; a_1 = 0.5; a_3 = 0.1; a_0 = 0$. Finally, ζ_t and M_t^g are AR(1) processes with persistence 0.95. The decision rules for real wage, output, interest rates, real balances and inflation, in terms of lagged real wages and the two shocks are

$$\begin{bmatrix} \widehat{w}_t \\ \widehat{y}_t \\ \widehat{i}_t \\ \widehat{M_t^{\dagger}} \\ \widehat{\Pi}_t \end{bmatrix} = \begin{bmatrix} 0.0012 \\ 0.5571 \\ 0.0416 \\ 0.1386 \\ 0.1050 \end{bmatrix} \begin{bmatrix} \widehat{w}_{t-1} \end{bmatrix} + \begin{bmatrix} 0.5823 & -0.0005 \\ 0.2756 & 0.0008 \\ 0.0128 & 0.9595 \\ 0.0427 & -0.1351 \\ -0.7812 & 0.0025 \end{bmatrix} \begin{bmatrix} \widehat{\zeta_t} \\ \widehat{M_t^g} \end{bmatrix}.$$

Two features of this approximate solution are worth commenting upon. First, there is little feedback from the state to the endogenous variables, except for output. This implies that the propagation properties of the model are limited. Second, monetary disturbances have little contemporaneous impact on all variables, except interest rates and real balances. These two observations imply that monetary disturbances have negligible real effects. This is confirmed by standard statistics. For example, technology shocks explain about 99 percent of the variance of output at the four years horizon and monetary shocks the rest. This model also misses the sign of few important contemporaneous correlations. For example, using linearly detrended US data the correlation between output and inflation is 0.35. For the model, the correlation is -0.89.

Exercise 2.33 *(Delivery lag) Suppose the representative agent maximizes $E_0 \sum_t \beta^t [\ln c_t - \vartheta_l N_t]$ subject to $c_t + i_t \leq \zeta_t K_t^{1-\eta} N_t^{\eta}$ and assume one period delivery lag, i.e. $K_{t+1} = (1-\delta)K_t + inv_{t-1}$. The Euler equation is $\beta E_t[c_{t+1}^{-1}(1-\eta)GDP_{t+1}K_{t+1}^{-1}] + (1-\delta)c_t^{-1} - \beta^{-1}c_{t-1}^{-1} = 0$. Log linearize the system and find a solution using K_t and $c_t^* = c_{t-1}$ as states.*

Vaughan's method, popularized by Blanchard and Kahn (1980) and King, Plosser and Rebelo (1987), takes a slightly different approach. First, using the state space representation for the (log)-linearized version of the model, it eliminates the expectation operator either assuming certainty equivalence or substituting expectations with actual values of the variables plus an expectational error. Second, it uses the law of motion of the exogenous variables, the linearized solution for the state variables and the costate (the Lagrangian multiplier) to create a system of first order difference equations (if the model delivers higher order dynamics, the dummy variable trick described in example 2.17 can be used to get the system in the required form). Third, it computes an eigenvalue-eigenvector decomposition on the matrix governing the dynamics of the system and divides the roots into explosive and stable ones. Then, the restrictions implied by the stability condition are used to derive the law of motion for the control (and the expectational error, if needed).

Suppose that the log-linearized system is $\Upsilon_t = \mathcal{A} E_t \Upsilon_{t+1}$ where $\Upsilon_t = [y_{1t}, y_{2t}, y_{3t}, y_{4t}]$, y_{2t} and y_{1t} are, as usual, the states and the controls, y_{4t} are the costates and y_{3t} are the shocks and partition $\Upsilon_t = [\Upsilon_{1t}, \Upsilon_{2t}]$. Let $\mathcal{A} = \mathcal{P}\mathcal{V}\mathcal{P}^{-1}$ be the eigenvalue-eigenvector decomposition of $\mathcal{A}$. Since the matrix $\mathcal{A}$ is symptletic, the eigenvalues come in reciprocal pairs when distinct. Let $\mathcal{V} = \text{diag}(\mathcal{V}_1, \mathcal{V}_1^{-1})$, where $\mathcal{V}_1$ is a matrix with eigenvalues greater than one in modulus and $\mathcal{P}^{-1} = \begin{bmatrix} \mathcal{P}_{11}^{-1} & \mathcal{P}_{12}^{-1} \\ \mathcal{P}_{21}^{-1} & \mathcal{P}_{22}^{-1} \end{bmatrix}$. Multiplying both sides by $\mathcal{A}^{-1}$, using certainty

equivalence and iterating forward we have

$$\left[\begin{array}{c}\Upsilon_{1t+j}\\ \Upsilon_{2t+j}\end{array}\right]=\mathcal{P}\left[\begin{array}{cc}\mathcal{V}_1^{-j} & 0\\ 0 & \mathcal{V}_1^{j}\end{array}\right]\left[\begin{array}{c}\mathcal{P}_{11}^{-1}\Upsilon_{1t}+\mathcal{P}_{12}^{-1}\Upsilon_{2t}\\ \mathcal{P}_{21}^{-1}\Upsilon_{1t}+\mathcal{P}_{22}^{-1}\Upsilon_{2t}\end{array}\right] \tag{2.54}$$

We want to solve (2.54) under the condition that Υ_{2t+j} goes to zero as $j\to\infty$, starting from some Υ_{20}. Since the components of $\mathcal{V}_1$ exceed unity, this is possible only if the terms multiplying $\mathcal{V}_1$ are zero. This implies $\Upsilon_{2t}=-(\mathcal{P}_{22}^{-1})^{-1}\mathcal{P}_{21}^{-1}\Upsilon_{1t}\equiv\mathcal{Q}\Upsilon_{1t}$ so that (2.54) is

$$\left[\begin{array}{c}\mathcal{Q}\Upsilon_{1t+j}\\ \Upsilon_{2t+j}\end{array}\right]=\left[\begin{array}{c}\mathcal{Q}\mathcal{P}_{11}\mathcal{V}_1^{-j}(\mathcal{P}_{11}^{-1}\Upsilon_{1t}+\mathcal{P}_{12}^{-1}\Upsilon_{2t})\\ \mathcal{P}_{21}\mathcal{V}_1^{-j}(\mathcal{P}_{11}^{-1}\Upsilon_{1t}+\mathcal{P}_{12}^{-1}\Upsilon_{2t})\end{array}\right] \tag{2.55}$$

which also implies $\mathcal{Q}=\mathcal{P}_{21}\mathcal{P}_{11}^{-1}$. Note that, for quadratic problem, the limit value of $\mathcal{Q}$ is the same as the limit of the Riccati equation (2.30).

Example 2.21 *The basic RBC model with labor-leisure choice, no habit, $G_t=T_t=T^y=0$, production function $f(K_t,N_t,\zeta_t)=\zeta_t K_t^{1-\eta}N_t^{\eta}$ and utility function $u(c_t,c_{t-1},N_t)=\ln c_t+\vartheta_N(1-N_t)$ when log linearized, delivers the representation $\Upsilon_t=\mathcal{A}_0^{-1}\mathcal{A}_1E_t\Upsilon_{t+1}$, where $\Upsilon_t=[\hat{c}_t,\hat{K}_t,\hat{N}_t,\hat{\zeta}_t]$ (since there is a one-to-one relationship between c_t,N_t and λ_t we can solve λ_t out of the system) where $\hat{}$ indicates percentage deviations from steady states and*

$$\mathcal{A}_0=\left[\begin{array}{cccc}1 & \eta-1 & 1-\eta & -1\\ -1 & 0 & 0 & 0\\ -(\frac{c}{K})^{ss} & (1-\eta)(\frac{N^{ss}}{K^{ss}})^{\eta}+(1-\delta) & \eta(\frac{N^{ss}}{K^{ss}})^{\eta} & (\frac{N^{ss}}{K^{ss}})^{\eta}\\ 0 & 0 & 0 & \rho\end{array}\right];$$

$$\mathcal{A}_1=\left[\begin{array}{cccc}0 & 0 & 0 & 0\\ -1 & -\beta\eta(1-\eta)(\frac{N^{ss}}{K^{ss}})^{\eta} & \beta\eta(1-\eta)(\frac{N^{ss}}{K^{ss}})^{\eta} & \beta(1-\eta)(\frac{N^{ss}}{K^{ss}})^{\eta}\\ 0 & 1 & 0 & 0\\ 0 & 0 & 0 & 1\end{array}\right].$$

Let $\mathcal{A}_0^{-1}\mathcal{A}_1=\mathcal{P}\mathcal{V}\mathcal{P}^{-1}$, where $\mathcal{P}$ is a matrix whose columns are the eigenvectors of $\mathcal{A}_0^{-1}\mathcal{A}_1$ and $\mathcal{V}$ contains, on the diagonal, the eigenvalues. Then

$$\mathcal{P}^{-1}\Upsilon_t\equiv\Upsilon_t^{\dagger}=\mathcal{V}E_t\Upsilon_{t+1}^{\dagger}\equiv\mathcal{V}E_t\mathcal{P}^{-1}\Upsilon_{t+1} \tag{2.56}$$

Since $\mathcal{V}$ is diagonal, there are four independent equations which can be solved forward, i.e.

$$\Upsilon_{it}^{\dagger}=v_iE_t\Upsilon_{i,t+\tau}^{\dagger}\qquad i=1,\ldots,4 \tag{2.57}$$

Since one of the conditions describes the law of motion of the technology shocks, one of the eigenvalues is ρ_{ζ}^{-1} (the inverse of the persistence of technology shocks). One other condition describes the intratemporal efficiency condition (see equation (2.14)): since this is a static relationship the eigenvalue corresponding to this equation is zero. The other two conditions, the Euler equation for capital accumulation (equation (2.13)) and the resource constraint (equation (2.5)) produce two eigenvalues: one above and one below one. The stable solution

is associated with the $v_i > 1$ since $\Upsilon_{it}^{\dagger} \to \infty$ for $v_i < 1$. Hence for (2.57) to hold for each t in the stable case, it must be that $\Upsilon_{it}^{\dagger} = 0$ for all $v_i < 1$.

Assuming $\beta = 0.99,\ \eta = 0.64,\ \delta = 0.025,\ \vartheta_n = 3$, the resulting steady states are $c^{ss} = 0.79;\ K^{ss} = 10.9;\ N^{ss} = 0.29,\ GDP^{ss} = 1.06$ and

$$\Upsilon_t^{\dagger} = \begin{bmatrix} 1.062 & 0 & 0 & 0 \\ 0 & 1.05 & 0 & 0 \\ 0 & 0 & 0.93 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} E_t \begin{bmatrix} -2.18 & -0.048 & 0.048 & 24.26 \\ 0 & 0 & 0 & 23.01 \\ -2.50 & 1.36 & 0.056 & 1.10 \\ -2.62 & 0.94 & -0.94 & 2.62 \end{bmatrix} \Upsilon_{t+1}^{\dagger}.$$

The second row has $v_2 = \rho_{\zeta}^{-1}$, the last one the intertemporal condition. The remaining two rows generate a saddle path. Setting the third and the forth rows to zero ($v_3, v_4 < 1$) we have $c_t = 0.54N_t + 0.02K_t + 0.44\zeta_t$ and $N_t = -2.78c_t + K_t + 2.78\zeta_t$. The law of motion of the capital stock can be read off the first equation: $K_{t+1} = -0.07c_t + 1.01K_t + 0.06N_t + 0.10\zeta_t$.

Exercise 2.34 *Suppose agents maximize the separable utility : $E_0 \sum_{t=0}^{\infty} u(c_t, 1 - N_t)$ by choices of consumption, hours and nominal money balances subject to the following three constraints:*

$$\begin{aligned} gdp_t &= \zeta_t N_t^{\eta} = G_t + c_t \\ c_t &= M_t / p_t \\ M_{t+1} &= (M_t - p_t c_t) + p_t(y_t - G_t) + M_t(\bar{M} + M_t^g) \end{aligned} \tag{2.58}$$

where ζ is a technology shock, G_t goverment expenditure, c_t consumption, M_t nominal balances and p_t prices. Here G_t, ζ_t, M_t^g are exogenous. Note that the third constraint describes the accumulation of money: $\bar{M}$ is a constant and M_t^g is a mean zero random variable.
i) Derive and log-linearize the first order condition of the problem. What are the states?
ii) Solve the linear system assuming that the growth rate of the exogenous variables (ζ_t, G_t, M_t^g) is an AR(1) process with common parameter ρ. Calculate the equilibrium expressions for inflation, output growth and real balances.
iii) Suppose you want to price the term structure of nominal bonds. Such bonds cost 1 unit of money at time t and give $1 + i_{t+\tau}$ units of money at time $t + \tau,\ \tau = 1, 2, \ldots$. Write the equilibrium conditions to price these bonds. Calculate the log-linear expression of the slope for the term structure between a bond with maturity $\tau \to \infty$ and a one period bond.
iv) Calculate the equilibrium pricing formula and the rate of return for stocks which costs p_t^s units of consumption at t and pays dividends $p_t^s sd_t$ which can be used for consumption only at $t + 1$ (Hint: the value of dividends at $t + 1$ is $p_t^s sd_t / p_{t+1}$). Calculate a log-linear expression for the equity premium (the difference between the nominal return on stocks and the nominal return on a 1 period bond).
v) Simulate the responses of the slope of term structure and of the equity premium to a unitary shock in the technology (ζ_t), in government expenditure (G_t) and in money growth (M_t^g). Is the pattern of responses economically sensible? Why?

Exercise 2.35 *(Pappa) Consider the sticky price model with the capital utilization setup analyzed in exercise 2.11 but without adjustment costs to capital. Log linearize the model*

and compute output responses to monetary shocks (still assume the monetary rule 2.22). How does the specification compare in terms of persistence and amplitude of real responses to the standard one, without capacity utilization, but with capital adjustment costs?

2.2.4 Second order approximations

First order (linear) approximation are fairly easy to construct, useful for a variety of purposes and accurate enough for fitting DSGE models to the data. However, first order approximations are insufficient, when evaluating welfare across policies that do not have direct effects on the deterministic steady state of the model, when analyzing asset pricing problems or when risk considerations become important. In some cases it may be enough to assume that nonlinearities, although important, are small in some sense (see e.g. Woodford (2002)). In general, one may want to have methods to solve second order (linear) system and producing locally accurate approximations to the dynamics of the model, without having to explicitly consider global (nonlinear) approximations.

Suppose the model has the form:

$$E_t \mathfrak{J}(y_{t+1}, y_t, \sigma\epsilon_{t+1}) = 0 \tag{2.59}$$

where $\mathfrak{J}$ is an $n \times 1$ vector of functions, y_t is an $n \times 1$ vector of endogenous variables and ϵ_t a $n_1 \times 1$ vector of shocks. Clearly, some components of (2.59) may be deterministic and others may be static. So far we have been concerned with the first order expansions of (2.59), i.e with the following system of equations

$$E_t[\mathfrak{J}_1(dy_{t+1} + \mathfrak{J}_2 dy_t + \mathfrak{J}_3 \sigma d\epsilon_{t+1}) = 0 \tag{2.60}$$

where dx_t is the deviation of x_t from some pivotal point. As we have seen, solutions to (2.60) are found positing a functional relationship $y_{t+1} = \mathfrak{J}^*(y_t, \sigma\epsilon_t, \sigma)$, linearly expanding it around the steady state $\mathfrak{J}^*(y^{ss}, 0, 0)$, substituting the linear expression in (2.60) and matching coefficients.

Here we are concerned with approximations of the form:

$$\begin{aligned} E_t[\mathfrak{J}_1(dy_{t+1} + \mathfrak{J}_2 dy_t + \mathfrak{J}_3 \sigma d\epsilon_{t+1} &+ \\ 0.5(\mathfrak{J}_{11} dy_{t+1} dy_{t+1} + \mathfrak{J}_{12} dy_{t+1} dy_t + \mathfrak{J}_{13} dy_{t+1} \sigma d\epsilon_{t+1} &+ \\ \mathfrak{J}_{22} dy_t dy_t + \mathfrak{J}_{23} dy_t \sigma d\epsilon_t + \mathfrak{J}_{33} \sigma^2 d\epsilon_{t+1} d\epsilon_{t+1}) &= 0 \end{aligned} \tag{2.61}$$

which are obtained from a second order Taylor expansion of (2.59). These differ from standard linearization with log-normal errors since a second order terms in dy_t, dy_{t+1} appear in the expression.

Since the second order terms enter linearly in the specification, solutions to (2.61) can also be obtained with the method of undetermined coefficients, assuming there exists a solution of the form $y_{t+1} = \mathfrak{J}^*(y_t, \sigma\epsilon_t, \sigma)$, taking a second order expansion of this guess around the steady states $\mathfrak{J}^*(y^{ss}, 0, 0)$, substituting the second order expansion for y_{t+1} into

(2.61) and matching coefficients. As shown by Schmitt Grohe and Uribe (2004), the problem can be sequentially solved, finding first the first order terms and then the second order ones.

Clearly, we need regularity conditions for the solution to exist and to have good properties. Kim, et. al. (2004) provide a set of necessary conditions. We first need that the solution implies that y_{t+1} remains in the stable manifold defined by $\mathfrak{H}(y_{t+1}, \sigma) = 0$ and satisfies $\{\mathfrak{H}(y_t, \sigma) = 0,\ \mathfrak{H}(y_{t+1}, \sigma) = 0$ almost surely and $\mathfrak{J}^1(y_{t+1}, y_t, \sigma\epsilon_{t+1}) = 0$ almost surely imply $E_t\mathfrak{J}^2(y_{t+1}, y_t, \sigma\epsilon_{t+1}) = 0\ \}$, where $\mathfrak{J} = (\mathfrak{J}^1, \mathfrak{J}_2)$. Second, we need $\mathfrak{H}(y_{t+1}, \sigma)$ to be continuous and twice differentiable in both its arguments Third, we need that the smallest unstable root of the first order system to exceed the square of its largest stable root. This last condition is automatically satisfied if the dividing line is represented by a root of 1.0, but, in general, one need to check that the roots of $\mathfrak{J}_1$ have this property.

Under these conditions, Kim et al, argue that the second order approximate solution to the dynamics of the model is accurate, in the sense that the error in the approximation converges in probability to zero at a more rapid rate than $||dy_t, \sigma||^2$, when $||dy_t, \sigma||^2 \to 0$. This claim does not depend on the almost sure boundness of the process for ϵ, which is violated when its distribution has unbounded support, or on the stationarity of the model. However, for non-stationary systems the n-step ahead accuracy deteriorates quicker than in the stationary case.

Example 2.22 *We consider a version of the two country model analyzed in example 2.6, where the population is the same in the two countries, the social planner equally weights the utility of the agents of the two country, there is no intermediate good sector, capital adjustment costs are zero and output is produced with capital only. The planner objective function is* $E_0 \sum_t \beta^t (\frac{c_{1t}^{1-\varphi}}{1-\varphi} + \frac{c_{2t}^{1-\varphi}}{1-\varphi}$*, the resource constraint is* $c_{1t} + c_{2t} + k_{1t+1} + k_{2t+1} - (1 - \delta)(k_{1t} + k_{2t}) = \zeta_{1t} k_{1t}^{1-\eta} + \zeta_{2t} k_{2t}^{1-\eta}$ *and* $\ln \zeta_{it},\ i = 1, 2$ *is assumed to be iid with mean zero and variance* σ^2*. Given the symmetry of the two countries, it must be the case that in equilibrium* $c_{1t} = c_{2t}$ *and that the Euler equations for capital accumulations in the two countries are identical. Letting* $\varphi = 2, \delta = 0.1, 1 - \eta = 0.3, \beta = 0.95$*, the steady state is* $(k_i, \zeta_i, c_i) = (2.62, 1, 00, 1.07), i = 1, 2$ *and a first order expansion of the policy function is*

$$\begin{bmatrix} k_{it+1} \end{bmatrix} = \begin{bmatrix} 0.444 & 0.444 & 0.216 & 0.216 \end{bmatrix} \begin{bmatrix} k_{1t} \\ k_{2t} \\ \zeta_{1t} \\ \zeta_{2t} \end{bmatrix} \tag{2.62}$$

A second order expansion of the policy function in country i is

$$k_{it+1} = \begin{bmatrix} 0.444 & 0.444 & 0.216 & 0.216 \end{bmatrix} \begin{bmatrix} k_{1t} \\ k_{2t} \\ \zeta_{1t} \\ \zeta_{2t} \end{bmatrix} - 0.83\sigma^2$$

$$+ \quad 0.5 \left[\begin{array}{cccc} k_{1t} & k_{2t} & \zeta_{1t} & \zeta_{2t} \end{array} \right] \left[\begin{array}{cccc} 0.22 & -0.18 & -0.02 & -0.08 \\ -0.18 & 0.22 & -0.08 & -0.02 \\ -0.02 & -0.08 & 0.17 & -0.04 \\ -0.08 & -0.02 & -0.04 & 0.17 \end{array} \right] \left[\begin{array}{c} k_{1t} \\ k_{2t} \\ \zeta_{1t} \\ \zeta_{2t} \end{array} \right] \quad (2.63)$$

Hence, apart for the quadratic terms in the states, first and second order solutions differ in the sense that the variance of the technology shock matter. In particular, when technology shocks are highly volatile, more consumption and less capital will be chosen with the second order approximation. Clearly, the variance of the shocks is irrelevant for the for the decision rules obtained with the first order approximation.

Exercise 2.36 *Consider the sticky price model whose log-linear approximation is described in exercise 2.30. Assuming that* $\vartheta = 0.5, \varphi = 2, \vartheta_M = 0.5, \zeta_p = 0.75, \beta = 0.99$ *find a first and second order expansions of the solution for* c_t, N_t, i_t, π_t*, assuming that there are only monetary shocks, which are iid with variance* σ^2*, that monetary policy is conducted using a rule of the form* $i_t = \pi_t^{a_3} M_t^g$ *and that* w_t *is equal to the marginal product of labor.*

2.2.5 Parametrizing expectations

The method of parametrizing expectations was suggested by Marcet (1989) and further developed by Marcet and Lorenzoni (1999)). With this approach the approximation is globally valid as opposed to valid only around a particular point as it is the case with quadratic, log-linear or second order approximations. Therefore, with such a method we can undertake experiments which are, e.g., far away from the steady state, unusual from the historical point of view or involve switches of steady states. The approach has two advantages: first, it can be used when inequality constraints are present. Second, it has a built-in mechanism that allows us to check whether a candidate solution satisfies the optimality conditions of the problem. Therefore, we can implicitly examine the accuracy of the approximation.

The essence of the method is simple. First, one approximates the expectational equations of the problem with a vector of functions $\hbar$, i.e. $E_t[h(y_{2t+1}, y_{2t}, y_{3t+1}, y_{3t})] \approx \hbar(\alpha, y_{2t}, y_{3t})$, where y_{2t} and y_{3t} are known at t and α is a vector of (nuisance) parameters. Polynomial, trigonometric, logistic or other simple functions which are known to have good approximation properties can be used. Second, one estimates α by minimizing the distance between $E_t[h(y_{2t+1}(\alpha), y_{2t}(\alpha), y_{3t+1}, y_{3t})]$ and $\hbar(\alpha, y_{2t}(\alpha), y_{3t})$, where $y_{2t}(\alpha)$ are simulated states from the approximate solution. Let α^* be the distance minimizer. Define

$$Q(\alpha, \alpha^*) = argmin_{\{\alpha^*\}} |E_t[h(y_{2t+1}(\alpha), y_{2t}(\alpha), y_{3t+1}, y_{3t})] - \hbar(\alpha^*, y_{2t}, y_{3t})|^q \quad (2.64)$$

some $q \geq 1$. The method then looks for a $\tilde{\alpha}$ such that $Q(\tilde{\alpha}, \tilde{\alpha}) = 0$.

Example 2.23 *Consider a basic RBC model with inelastic labor supply, utility given by* $u(c_t) = \frac{c_t^{1-\varphi}}{1-\varphi}$*, where* φ *is a parameter, budget constraint* $c_t + K_{t+1} + G_t = (1 - T^y)\zeta_t K_t^{1-\eta} +$

$(1-\delta)K_t + T_t$ *and let* $(\ln\zeta_t, \ln G_t)$ *be AR processes with persistence* $(\rho_\zeta,\ \rho_g)$ *and unit variance. The expectational (Euler) equation is*

$$c_t^{-\varphi} = \beta E_t[c_{t+1}^{-\varphi}((1-T^y)\zeta_{t+1}(1-\eta)K_{t+1}^{-\eta} + (1-\delta))] \tag{2.65}$$

where β *is the rate of time preferences. We wish to approximate the expression on the right hand side of (2.65) with a function* $\hbar(K_t, \zeta_t, G_t, \alpha)$*, where* α *is a set of parameters. Then the parametrizing expectation algorithm works as follows:*

Algorithm 2.3

1) *Select* $(\varphi, T^y, \delta, \rho_\zeta, \rho_g, \eta, \beta)$*. Generate* (ζ_t, G_t)*,* $t = 1, \ldots T$*, choose an initial* α^0*.*

2) *Given a functional form for* $\hbar$ *calculate* $c_t(\alpha^0)$ *from (2.65) with* $\hbar(\alpha^0,\ k_t,\ \zeta_t,\ G_t)$*, in place of* $\beta E_t[c_{t+1}^{-\varphi}((1-T^y)\zeta_{t+1}(1-\eta)K_{t+1}^{-\eta} + (1-\delta))]$ *and* $K_{t+1}(\alpha^0)$ *from the resource constraint. Do this for every* t*. This produces a time series for* $c_t(\alpha^0)$ *and* $K_{t+1}(\alpha^0)$*.*

3) *Run a nonlinear regression using simulated* $c_t(\alpha^0), K_{t+1}(\alpha^0)$ *of* $\hbar(\alpha, K_t(\alpha^0), \zeta_t, G_t)$ *on* $\beta c_{t+1}(\alpha^0)^{-\varphi}((1-T^y)\zeta_{t+1}(1-\eta)K_{t+1}(\alpha^0)^{-\eta} + (1-\delta)$*. Call the resulting nonlinear estimator* α^{0*} *and with this* α^{0*} *construct* $Q(\alpha^0, \alpha^{0,*})$*.*

4) *Set* $\alpha^1 = (1-\varrho)\alpha^0 + \varrho Q(\alpha^0, \alpha^{0*})$*, where* $\varrho \in (0, 1]$*.*

5) *Repeat steps 2)-4) until until* $Q(\alpha^{*L-1}, \alpha^{*L}) \approx 0$ *or* $|\alpha^L - \alpha^{L-1}| \leq \iota$*, or both,* ι *small.*

6) *Use another* $\hbar$ *function and repeat steps 2)-5).*

When convergence is achieved $\hbar(\alpha^*, K_t, \zeta_t, G_t)$ *is the required approximating function. Since the method does not specify how to choose* $\hbar$*, it is typical to start with a simple function (a first order polynomial or a trigonometric function) and then in 6) check the robustness of the solution using more complex functions (e.g a higher order polynomial).*

For the model of this example, setting $\varphi = 2, T^y = 0.15, \delta = 0.1, \rho_g = \rho_\zeta = 0.95, \eta = 0.66, \beta = 0.99, q = 2$ *and choosing* $\hbar = \exp(\ln\alpha_1 + \alpha_2 \ln K_t + \alpha_3 \ln\zeta_t + \alpha_4 \ln G_t)$*, 100 iterations of the above algorithm led to the following optimal approximating values* $\alpha_1 = -0.0780, \alpha_2 = 0.0008, \alpha_3 = 0.0306, \alpha_4 = 0.007$ *and with these values* $Q(\alpha^{*L-1}, \alpha^{*L}) = 0.000008$*.*

We show how to apply the method when inequality constraints are present next.

Example 2.24 *Consider a small open economy which finances current account deficits issuing one period nominal bonds. Assume that there is a borrowing constraint* $\bar{B}$ *so that* $B_t - \bar{B} < 0$*. The Euler equation for debt accumulation is*

$$c_t^{-\varphi} - \beta E_t[c_{t+1}^{-\varphi}(1+r_t) - \lambda_{t+1})] = 0 \tag{2.66}$$

where r_t *is the exogenous world real rate,* λ_t *the Lagrangian multiplier on the borrowing constraint and the Kuhn-Tucker conditions is* $\lambda_t(B_t - \bar{B}) = 0$*. To find a solution use*

$0 = c_t^{-\varphi} - \beta\hbar(\alpha, r_t, \lambda_t, c_t)$ *and* $\lambda_t(B_t - \bar{B}) = 0$ *and calculate* c_t *and* B_t*, assuming* $\lambda_t = 0$*. If* $B_t > \bar{B}$ *set* $B_t = \bar{B}$*, find* λ *from the first equation and* c_t *from the budget constraint. Do this for every* t*; find* α^{0*}*; generate* α^1 *and repeat until convergence. Hence,* λ_t *is treated as an additional variable, to be solved for in the model.*

Exercise 2.37 *Suppose in the model of example 2.23 that* $u(c_t, c_{t-1}, N_t) = \frac{(c_t - \gamma c_{t-1})^{1-\varphi}}{1-\varphi}$*,* $T_t = T^y = 0$*. Provide a parametrized expectation algorithm to solve this model (Hint: there are two state variables in the Euler equation).*

Exercise 2.38 *(CIA with taxes) Consider a model where agents maximize a separable utility function of the form:* $E_0 \sum_{t=0}^{\infty} \beta^t(\vartheta_c \ln(c_{1t}) + (1 - \vartheta_c)\ln(c_{2t}) - \vartheta_N(1 - N_t))$ *by choices of consumption of cash and credit goods, leisure, nominal money balances and investments,* $0 < \beta < 1$*. Suppose that the household is endowed with* K_0 *units of capital and one unit of time. The household receives income from capital and labor which is used to finance consumption purchases, investments and holdings of money and government bonds.* c_{1t} *is the cash-good and needs to be purchased with money,* c_{2t} *is the credit good. Output is produced with capital and labor by a single competitive firm with constant returns to scale technology and* $1 - \eta$ *is the share of capital. In addition, the government finances a stochastic flow of expenditure by issuing currency, taxing labor income with a marginal tax rate* T_t^y *and issuing nominal bonds, which pay an interest rate* i_t*. Assume that money supply evolves according to* $\ln M_{t+1}^s = \ln M_t^s + \ln M_t^g$*. Suppose agents start at time* t *with holdings of money* M_t *and bonds* B_t*. Assume that all the uncertainty is resolved at the beginning of each* t*.*
i) Write down the optimization problem mentioning the states and the constraints and calculate the first order conditions (Hint: you need to make the economy stationary).
ii) Solve the model parametrizing the expectations and using a first order polynomial.
iii) Describe the effects of an iid shock in T_t^y *on real variables, prices and interest rates, when* B_t *adjusts to satisfy the government budget constraint. Would your answer change if you keep* B_t *fixed and let instead* G_t *change to satisfy the government constraint?*

Example 2.25 *(Marcet and Den Haan) The method of parametrizing expectations has a built-in mechanism to check the accuracy of the approximation. In fact, whenever the approximation is appropriate, the simulated time series must satisfy the Euler equation. As we will describe in more details in chapter 5, this implies that if* $\tilde{\alpha}$ *solves* $Q(\tilde{\alpha}, \tilde{\alpha}) = 0$ *then* $Q(\tilde{\alpha}, \tilde{\alpha}) \otimes h(z_t) = 0$*, where* z_t *is any variable in the information set at time* t*;* h *is a* $q \times 1$ *vector of continuous differential functions. Under regularity conditions, when* T *is large,* $\mathfrak{S} = T \times (\frac{1}{T}\sum_t (Q_t \otimes h(z_t)))' W_T (\frac{1}{T}\sum_t (Q_t \otimes h(z_t))) \to \chi^2(\nu)$*, where* Q_t *is the sample counterpart of* Q*,* ν *is equal to the dimension of the Euler conditions times the dimension of* h *and* $W_T \xrightarrow{P} W$ *is a weighting matrix. For the example 2.23, the first order approximation is accurate since* $\mathfrak{S}$ *has a p-value of 0.36, when two lags of consumption are used as* z_t*.*

While useful in a variety of problems, the parametrizing expectations approach has two important drawbacks. First, the iterations defined by algorithm 2.3 may led nowhere since the fixed point problem does not define a contraction operator. In other words, there is

no guarantee that the distances between the actual and approximating function will get smaller as the number of iterations grows. Second, the method relies on the sufficiency of the Euler equations. Hence, if the utility function is not strictly concave, the solution that the algorithm delivers may be inappropriate.

2.2.6 A Comparison of methods

There exists some literature comparing various approximation approaches. For example, the special issue of the Journal of Business and Economic Statistics of July 1991 shows how various methods perform in approximating the decision rules of a particular version of the one sector growth model for which analytic solutions are available. Some additional evidence is in Ruge-Murcia (2002) and Fernandez-Villaverde and Rubio-Ramirez (2003). In general, little is known about the properties of various methods in specific applications. Experience suggests that even for models possessing simple structures (i.e. models without habit, adjustment costs of investment, etc.), simulated series may display somewhat different dynamics depending on the approximation used. For more complicated models no evidence is available. Therefore, caution should be employed in interpreting the results obtained approximating models with any of the methods described in this chapter.

Exercise 2.39 *(Growth with corruption) Consider a representative agent who maximizes* $E_0 \sum_t \beta_t \frac{c_t^{1-\varphi}}{1-\varphi}$ *by choices of consumption* c_t*, capital* K_{t+1} *and bribes* br_t *subject to*

$$c_t + K_{t+1} = (1 - T_t^y) N_t w_t + r_t K_t - br_t + (1-\delta) K_t \tag{2.67}$$

$$T_t^y = T_t^e (1 - a \ln br_t) + T_0^y \tag{2.68}$$

where w_t *is the real wage,* T_t^y *is the income tax rate,* T_t^e *is an exogenously given tax rate,* T_0^y *is the part of the tax rate which is unchanged by bribes, and* (φ, a, δ) *are parameters. The technology is owned by the firm and it is given by* $f(K_t, N_t, \zeta_t, K_t^G) = \zeta_t K_t^{1-\eta} N_t^{\eta} (K_t^G)^{\aleph}$*, where* $\aleph \geq 0$*,* K_t *is the capital stock and* N_t *the hours. Government capital* K_t^G *evolves according to* $K_{t+1}^G = (1-\delta) K_t^G + N_t w_t T_t^y$*. The resource constraint for the economy is* $c_t + K_{t+1} + K_{t+1}^G + br_t = f(K_t, N_t, \zeta_t) + (1-\delta)(K_t + K_t^G)$ *and* (ζ_t, T_t^e) *are independent AR(1) processes, with persistence* (ρ_ζ, ρ_e) *and variances* $(\sigma_\zeta^2, \sigma_e^2)$*.*
i) Define a competitive equilibrium and compute the first order conditions.
ii) Assume $\varphi = 2$*,* $a = 0.03$*,* $\beta = 0.96$*,* $\delta = 0.10$*,* $\rho_e = \rho_\zeta = 0.95$ *and set* $\sigma_\zeta^2 = \sigma_e^2 = 1$*. Take a quadratic approximation of the utility and find the decision rule for the variables of interest.*
iii) Assume that (ζ_t, T_t^e) *and the capital stock can take only two values (say, high and low). Solve the model discretizing the state space (Hint: use the fact that shocks are independent and the values of the AR parameter to construct the transition matrix for the shocks).*
iv) Take a log-linear approximation of the first order conditions of the problem. Solve the model using a first order linear approximation method.
v) Use the parametrized expectations method with a first order power function to find a global solution.

vi) Compare the time series properties for consumption, investment and bribes across methods. Are they different? In what?

Exercise 2.40 *(Transmission with borrowing constraints) Consider an economy where preferences are described by* $u(c_t, c_{t-1}, N_t) = \frac{(c_t^{\vartheta} N_t^{1-\vartheta})^{1-\varphi}}{1-\varphi}$ *and accumulates capital according to* $K_{t+1} = (1-\delta)K_t + inv_t$*, where* δ *is the depreciation rate. Assume that the production function is Cobb-Douglas in hours* (N_t)*, capital* (K_t)*, and land* (La_t) *and of the form* $GDP_t = \zeta_t K_t^{\eta_k} N_t^{\eta_N} La_t^{\eta_L}$*. Suppose individual agents have the ability to borrow and trade land and that their budget constraint is* $c_t + K_{t+1} + B_{t+1} + p_t^L La_{t+1} \leq GDP_t + (1-\delta)K_t + (1+r_t^B)B_t + p_t^L La_t$*, where* B_t *are bond holdings, and suppose that there is a borrowing constraint of the form* $p_t^L La_t - B_{t+1} \geq 0$*, where* p_t^i *is the price of land in terms of consumption goods*
(i) Show that in the steady state the borrowing constraint is binding if $\frac{GDP^{ss}}{K^{ss}} + (1-\delta) > (1 + r_t^B)$*. Give conditions which insure that the constraint is always binding.*
(ii) Describe the dynamics of output following a technology shock when: (a) the borrowing constraint never binds, (b) the borrowing constraint always binds, (c) the borrowing constraints binds at some t*. (Hint: Use an approximation method which allows the comparison across cases).*
(iii) Is it true that the presence of (collateralized) borrowing constraints amplifies and stretches over time the real effects of technology shocks?

Chapter 3: Extracting and Measuring Cyclical Information

Most of the models considered in chapter 2 are designed to explain or replicate cyclical features of the actual data. Unfortunately, most economic time series display trends or marked growth patterns so that it is not immediately obvious what the cyclical properties of the data are. This chapter is concerned with the process of obtaining cyclical information from the actual data and with the problem of efficiently and meaningfully summarizing it.

Cyclical information can be obtained in many ways. For example, Burns and Mitchell (1943) and the traditional cycle dating literature look at turning points of a reference series to extract this information. Following Lucas (1977), in macroeconomics it is however more common to obtain the cyclical information first eliminating a permanent component (the "trend") of the data, typically thought to be unrelated with those features that business cycle models are interested in explaining, and then computing second moments for the residuals (the " cycle" or better the "growth cycle"). In practice, since trends and cycles are unobservable, assumptions are needed to split observable series into components. Many assumptions can be make and it is impossible to formally choose among alternatives with a finite stretch of data. This means that standard criteria, such as lack of empirical relevance or statistical optimality, cannot be used.

The picture is further complicated by the fact that the literature has interchangeably used the terms detrending and filtering for the process of extracting growth cycles, even though the two concepts are distinct. Detrending should be intended as the process of making economic series (covariance) stationary. Detrending is necessary if one wants to compute functions of second moments of the data, which may not exist if a time series is e.g. a random walk, to estimate the parameters of the model with several of the procedures described in this book, but unnecessary for other purposes, such as dating turning points, measuring amplitudes or the responses to behavioral shocks. That is, certain cyclical information can be directly obtained from the raw data without any need of detrending.

The term filtering has a much broader applicability. As we have seen in chapter 1, filters are operators which carve out particular frequencies of the spectrum. One can build filters to eliminate very low frequency movements, to emphasize the variability in a particular frequency range, to smooth out high frequency idiosyncratic movements, or to mitigate the effect of measurement errors. Filtering is unnecessary when comparing the cyclical behavior

of the data to those of models, but it may help to bring out more clearly the differences between the two which are of most interest. In particular, since variability at frequencies corresponding to cycles of 6 to 24-32 quarters is considered of crucial economic importance - because business cycles reported, e.g. by the NBER and the CEPR, have periodicity which is, approximately, in this range - filtering may facilitate the comparison.

An important source of confusion emerges when time series econometricians and applied macroeconomists attempt to communicate the results of their studies since the formers attempt to isolate "periodic" components in growth cycles, that is, that is components which are representable with some form of sine and cosine functions and that show up as peaks in the spectral density in a particular frequency band. The latters, on the other hand, often consider business cycle phenomena the presence of serially correlate movements in the growth cycle (see e.g. Long and Plosser (1983)). Therefore, while they are satisfied when the growth cycle produced by their model is, for example, an AR(1) process, time series econometricans often use the argument that AR(1) have no peaks at business cycle frequencies and therefore there is no business cycle to speak of.

A final problem arises because some economic models feature shocks which have both transitory (short run) and permanent (long run) effects, in which case decompositions which assume that the two phenomena are separate are wrong; or because the effects they describe are not necessarily linked to statistical permanent/growth cycle decompositions. Monetary disturbances are a classic example of shocks having transitory effects on real variables (meaning not specifically located at any frequencies or of a particular periodicity) and permanent effect on nominal ones (meaning, in this case full, pass-though in the long run). Because the link between economic theory and empirical practice is embryonic, and because there is little consensus on the type of economic model one should use to guide the decomposition (which shocks are permanent?, which are transitory?, which dominates?, etc.), it is also hard to use economic theory to guide the decomposition.

Because traditional procedures displayed both conceptual and practical problems, new approaches have appeared over the last 20 years. We describe a subset of these methods, characterize their properties, discuss their relative merits, and highlight possible distortions that may appear when using them in comparing the output of a DSGE model and the data. We consider both univariate and multivariate methods and categorize decompositions into three somewhat arbitrary classes: statistical methods, economic methods and hybrid methods. In the first class we include procedures which have a statistical or a probabilistic justification. They use time series assumptions on the observable or the trend to measure the cycle. In the second class, extraction procedures are dictated by economic theory. Here, the cycles we obtain have relevance only to the extent that the model is a valid approximation to the data generating process (DGP). In the third class we include procedures which are statistical in nature but have an economic justification of some sort.

Throughout this chapter we denote the logarithm of the observables by y_t, their growth rate by $\Delta y_t = (1-\ell)y_t$, the trend (permanent component) by y_t^x and the cycle by y_t^c. Also, we use the convention that a variable is integrated of order d_0 $[I(d_0)]$ if it is (covariance) stationary after d_0-differencing.

3.1 Statistical Decompositions

3.1.1 Traditional methods

Traditionally, the trend of a series was taken to be deterministic and the cyclical component was measured as the residual of a regression of y_t on polynomials in time. That is, $y_t = y^x + y_t^c$; $y_t^x = a_0 + \sum_j a_j t^j$ and $\text{corr}(y_t^x, y_t^c) = 0$ so that $\hat{y}_t^c = y_t - \hat{a}_0 - \sum_j \hat{a}_j t^j$ and $\hat{a}_j, j = 0, 1, 2, \ldots$ are estimates of a_j. While the trend in such a setup can easily be estimated with least square methods, the specification is unsatisfactory in two senses. First, since the trend is deterministic, it can be perfectly predicted arbitrarily far into the future. Second, the growth rate of y_t cannot accelerate or decelerate, contradicting the evidence of a number of macroeconomic time series in many countries since WWII. This latter problem can be partially eliminated if we allow for structural breaks at preselected points.

Example 3.1 *Suppose* $y_t = y^x + y_t^c$, $y_t^x = a_{10} + \sum_j a_{1j} t^j$ *for* $t < t_1$ *and* $y_t^x = a_{20} + \sum_j a_{2j} t^j$ *for* $t \geq t_1$, $a_{1j} \neq a_{2j}$, *some* $j \geq 0$ *and let* $corr(y_t^x, y_t^c) = 0\ \forall t$. *Then* $\hat{y}_t^c = y_t - \hat{a}_{10} - \sum_j \hat{a}_{1j} t^j$ *for* $t < t_1$; $\hat{y}_t^c = y_t - \hat{a}_{20} - \sum_j \hat{a}_{2j} t^j$ *for* $t \geq t_1$. *Multiple breaks at known dates can be similarly obtained.*

The traditional alternative to linear/segmented trend specifications is to assume that the growth rate of y_t captures the cyclical properties of the data. Here $\Delta y_t = y_t^c$ and therefore $y_t^x = y_{t-1}$. This approach is also simple but has several disadvantages. First, the time plot of y_t^c does not visually conform to the idea one has of cyclical fluctuations. Second, y_t^c does not necessarily have zero mean. Third, somewhat counterintuitively, the variance of y_t^x may be very large.

Exercise 3.1 *Let* $\Delta y_t = y_t^c$. *Show that* $E(\Delta y_t^c \Delta y_{t-\tau}^c) = 2ACF_y(\tau) - ACF_y(\tau - 1) - ACF_y(\tau + 1)$, *where* $ACF_y(\tau) = cov(y_t, y_{t-\tau})$. *What can you say about the autocorrelation properties of* y_t^c *if* $y_t = \rho y_{t-1} + e_t,\ e_t \sim iid(0, 1), 0 < \rho < 1$*?*

3.1.2 Beveridge-Nelson (BN) decomposition

Beveridge and Nelson (1981) also assume that y_t is integrated of order one but define the trend as the conditional mean of the predictive distribution for future y_t's. Then the cyclical component is the forecastable momentum in y_t at each t. Let y_t be represented as:

$$\Delta y_t = \bar{y} + D(\ell) e_t \quad e_t \sim\ iid\ (0, \sigma_e^2) \tag{3.1}$$

where $D(\ell) = 1 + D_1 \ell + D_2 \ell^2 + \ldots$, $\bar{y}$ is a constant, and the roots of $D(\ell)$ lie on or outside the complex unit circle. Let the forecast of $y_{t+\tau}$ based on time t information be $y_t(\tau) \equiv E(y_{t+\tau} | y_t, y_{t-1}, \ldots, y_0) = y_t + E[\Delta y_{t+1} + \ldots + \Delta y_{t+\tau} | \Delta y_t, \ldots, \Delta y_0] \equiv y_t + \sum_{j=1}^{\tau} \widehat{\Delta y}_t(j)$. Using (3.1), $\widehat{\Delta y}_t(j) = \bar{y} + D_j e_t + D_{j+1} e_{t-1} + \ldots$ so that $y_t(\tau) = y_t + \tau \bar{y} + (\sum_{j=1}^{\tau} D_j) e_t + (\sum_{j=2}^{\tau+1} D_j) e_{t-1} + \ldots$. Let y_t^x be the time t forecast of $y_{t+\tau}$, adjusted for its mean rate of

change, i.e. $y_t^x \equiv y_t(\tau) - \tau\bar{y}$. For τ large, $y_t(\tau)$ is approximately constant and y_t^x is the value the series would have taken if it were on its long-run path. Hence

$$\lim_{\tau\to\infty} y_t^x = \lim_{\tau\to\infty}[y_t + (\sum_{j=1}^{\tau} D_j)e_t + (\sum_{j=2}^{\tau} D_j)e_{t-1} + \ldots] = y_{t-1}^x + \bar{y} + (\sum_{j=0}^{\infty} D_j)e_t \quad (3.2)$$

where the second equality comes from the fact that $(\sum_{j=1}^{\infty} D_j)e_t$ is a white noise and that $y_t^x - y_{t-1}^x = y_t - y_{t-1} + (\sum_{j=1}^{\infty} D_j)e_t - \sum_{j=1}^{\infty} D_j e_{t-j}$. Hence the trend is a random walk and the cyclical component is $y_t^c = y_t - y_t^x = -\sum_{j=0}^{\infty}(\sum_{i=j+1}^{\infty} D_i)e_{t-j}$.

One advantage of the BN decomposition over traditional approaches is that it produces a decomposition without any assumptions on the structure of the components or on their correlation. In fact, since it uses a forecast based definition of the trend, it does not need additional identifying restrictions to become operative.

Example 3.2 *(Pagan and Harding) Suppose* $\Delta y_t - \bar{y} = \rho(\Delta y_{t-1} - \bar{y}) + e_t,\ \rho < 1, e_t \sim iid(0, \sigma_e^2)$. *Then* $y_t^x = y_t + \frac{\rho}{1-\rho}(\Delta y_t - \bar{y})$ *and* $y_t^c = -\frac{\rho}{1-\rho}(\Delta y_t - \bar{y})$. *Since* ρ *is a constant, the properties of* y_t^c *and* Δy_t *are similar. Hence, for simple AR(1) processes, using the BN and growth rate decompositions give cycles with similar correlation properties.*

Several interesting features of the BN decomposition should be noted. First, since the two components are driven by the same shock, trend and cycle are perfectly correlated. Second, since estimates of D_j and forecasts $\widehat{\Delta y}_t(j)$ are typically obtained from ARIMA models, the standard identification problems of ARIMA specifications plague this method. Third, since long run forecasts of Δy_t are based on past values of y_t only, trend estimates may be very imprecise and estimates of var (Δy_t^c)/var (Δy_t^x) arbitrarily small. Finally, since innovations in y_t^x are $e_t^x = (\sum_{j=0}^{\infty} D_j)e_t$, the variability of the innovations in the trend may be larger than the variability of the innovations in the series.

Example 3.3 *Suppose* $y_t = y_{t-1} + \bar{y} + e_t + D_1 e_{t-1}$ *with* $0 < |D_1| < 1$ *and* $e_t \sim$ *iid* $(0, \sigma_e^2)$. *Note that if* D_1 *is positive,* Δy_t *is positively correlated. Then* $\lim_{\tau\to\infty} \Delta y_t^x = \bar{y} + (1+D_1)e_t = \bar{y} + e_t^x$ *and* $y_t^c = -D_1 e_t$. *Here* $var(e_t^x) > var(e_t)$ *and* y_t^c *is a white noise. In general, if* $D_j > 0\ \forall j \geq 1$, $var(e_t^x) > var(e_t)$. *Note that, if* $D_1 = 0$, $\Delta y_t^x = e_t$ *and* $y_t^c = 0$. *Hence, the presence of AR components are necessary to have serially correlated cycles and the decomposition correctly recognizes that if the AR components are missing, the cycle is either a white noise or inexistent.*

Exercise 3.2 *(Coddington and Winters): Show that if* $\Delta y_t = \bar{y} + \frac{D_2(\ell)}{D_1(\ell)}e_t$, $e_t \sim$ *iid* $(0, \sigma_e^2)$, y_t^x *satisfies* $y_t^x = y_{t-1}^x + \bar{y} + \frac{(1-\sum_{j=1}^{d_2} D_{2j})}{(1-\sum_{j=1}^{d_1} D_{1j})}e_t$, *where* d_1 *and* d_2 *are the lengths of the polynomial* $D_1(\ell)$ *and* $D_2(\ell)$. *and Suggest a way to recursively estimate* y_t^x.

Exercise 3.3 *Suppose* $y_t = (1-A)y_{t-1} + Ay_{t-2} + e_t,\ \ e_t \sim$ *iid* $(0, \sigma_e^2)$. *Find* y_t^x *and* y_t^c.

Extending the BN decomposition to multivariate frameworks is straightforward (see e.g. Evans and Reichlin (1994)). Let $y_t = [\Delta y_{1t}, y_{2t}]$ be an $(m \times 1)$ vector of stationary processes, where y_{1t} are I(1) variables and y_{2t} are (covariance) stationary; assume $y_t = \bar{y} + D(\ell)e_t$, where $e_t \sim \ iid\ (0, \Sigma_e)$ and (i) $D_0 = I$; (ii) the roots of $det(D(\ell))$ are on or outside the complex unit circle; (iii) $D_1(1) \neq 0$, where $D_1(\ell)$ is the matrix formed with the first m_1 rows of $D(\ell)$. Condition (i) is a simple normalization, condition (ii) insures that $D(\ell)$ is invertible so that e_t are the innovations in y_t; the latter condition insures the existence of at least one stochastic trend. Note that for $m_1 = m$ there are m stochastic trends and that $m_1 \neq 0$ is necessary for the decomposition to be meaningful. Then the multivariate Beveridge and Nelson decomposition is

$$\begin{pmatrix} \Delta y_{1t} \\ \Delta y_{2t} \end{pmatrix} = \begin{pmatrix} \bar{y}_1 \\ 0 \end{pmatrix} + \begin{pmatrix} D_1(1) \\ 0 \end{pmatrix} e_t + \begin{pmatrix} (1-\ell)D_1^{\dagger}(\ell) \\ (1-\ell)D_2^{\dagger}(\ell) \end{pmatrix} e_t \tag{3.3}$$

where $D_1^{\dagger}(\ell) \equiv \frac{D_1(\ell)-D_1(1)}{1-\ell}$, $D_2^{\dagger}(\ell) \equiv \frac{D_2(\ell)}{1-\ell}$, $\text{rank}[D_1(1)] \leq m_1$ and $y_t^x = y_{t-1}^x + [\bar{y}_1 + D_1(1)e_t, 0]'$ is the trend (permanent component) of y_t.

Example 3.4 *It is easy to verify that (3.3) is consistent with a univariate decomposition. Let the first component of y_{1t} be y_{1t}^1. Then $y_{1t}^{x1} = \lim_{\tau\to\infty}[E_t y_{1t+\tau}^1 - \tau\bar{y}^1]$, where $\bar{y}^1$ is the first element of $\bar{y}$. Hence $\Delta y_{1t}^{x1} = \bar{y}^1 + D_1^1(1)e_t$ where D_1^1 is the first row of D_1 and $\Delta y_{1t}^{c1} = (1-\ell)D_1^{1\dagger}(\ell)e_t$, where $D_1^{1\dagger}(\ell) = \frac{D_1^1(\ell)-D_1^1(1)}{1-\ell}$, and $y_{1t}^{c1} = \sum_j(\sum_{i=j+1} D_{1i}^1)e_{t-j}$.*

Exercise 3.4 *Consider a system with output, prices, interest rates and money and US quarterly data. Suggest a way to estimate the two components of a multivariate BN decomposition (Hint: Specify a VAR of the form $A(\ell)y_t = \bar{y} + e_t$ and find out which variables are in y_{1t} and which in y_{2t}).*

Three properties of multivariate BN decompositions should be noted. First, $\text{var}(\Delta y_t^{x1}) = D_1(1)\Sigma_e D_1(1)' = \mathcal{S}_{\Delta y_1}(0)$. That is, the variance of the permanent component is equal to the spectral density of Δy_{1t} at frequency $\omega = 0$. Hence, the spectral density of Δy_{1t}^c at $\omega = 0$ is zero. Second, permanent and transitory components can be obtained without identifying meaningful economic shocks. Third, the variance the cycle depends on the variables used to forecast y_t. In fact, $var(\Delta y_{1t}^c) = var(\Delta y_{1t}) + var(\Delta y_{1t}^x) + cov(\Delta y_{1t}, \Delta y_{1t}^x)$ which implies, after a few manipulations, $\frac{var(\Delta y_{1t}^c)}{var(\Delta y_{1t}^x)} \geq \frac{var(\Delta y_{1t})}{var(\Delta y_{1t}^x)} + 1 - 2\sqrt{\frac{var(\Delta y_{1t}|\mathcal{F}_{t-1})}{var(\Delta y_{1t}^x)}}$, where $\mathcal{F}_{t-1}$ is the info set available at t. Since $var(\Delta y_{1t}|\mathcal{F}_{t-1}^1) \leq var(\Delta y_{1t}|\mathcal{F}_{t-1}^2)$ if $\mathcal{F}_{t-1}^2 \subset \mathcal{F}_{t-1}^1$, it is possible to increase the variability of the (difference of the) cycle relative to the variability of the (difference of the) trend by enlarging the set of variables included in y_t (adding irrelevant variables does not help). Hence, the magnitude of $\text{var}(\Delta y_{1t}^c) - \text{var}(\Delta y_{1t}^x)$ in univariate and multivariate decompositions could be dramatically different.

Example 3.5 *Let y_t be a 2×1 vector and let the first element be $\Delta y_{1t} = e_{1t} + D_{11}e_{1t-1} + D_{21}e_{2t-1}$. Here var (Δy_{1t}^c) - var $(\Delta y_{1t}^x) = [D_{11}^2 - 1 - 2D_{11}]\sigma_{e_1}^2 + D_{21}^2\sigma_{e_2}^2 \geq [D_{11}^2 - 1 - 2D_{11}]\sigma_{e_1}^2 =$*

var (Δy_t^c) *- var* (Δy_t^x) *obtained from a univariate model. Note that, if* D_{21} *is large enough or if* $\sigma_{e_2}^2 >> \sigma_{e_1}^2$, *var* (Δy_{1t}^c) *- var* (Δy_{1t}^x) *could be positive even if var* (Δy_t^c) *- var* (Δy_t^x) *is negative. Clearly, complicated patterns may arise when* y_t *is a large scale VAR.*

Exercise 3.5 *Show that if* Δy_t *is positively correlated at all lags: (i)* $var(\Delta y_t) < var(\Delta y_t^x)$*; (ii)* $cov(\Delta y_t^c, \Delta y_t^x) < 0$*; (iii)* $cov(\Delta y_t, y_t^c) = -\sum_{\tau=1}^{\infty} ACF_{\Delta y_t}(\tau) < 0$.

As in the univariate case, the properties of estimated multivariate BN decompositions depend on a number of auxiliary assumptions e.g., the lag length of the model, the number of cointegrating relationships, etc. It is therefore important to carefully monitor the sensitivity of the results and the quality of the estimates to alterations in these assumptions.

Exercise 3.6 *Consider a bivariate VAR(1) where both variables are I(1). Show how to compute a multivariate BN decomposition in this case. Repeat the exercise when one variable is I(2) and one is I(1).*

3.1.3 Unobservable Components (UC) decompositions

Unobservable components decompositions are popular in the time series literature since cycle estimates obtained enjoy certain optimality properties (see e.g. Harvey (1989)). UC specifications are generally preferred to ARIMA representations for obtaining cyclical components for two reasons. First, there is no guarantee that an ARIMA model identified with standard methods will have those features that a series is postulated to exhibit (e.g. a cycle of the BN type requires the identification of an AR component). Second, the ARIMA(0,1,1) model favoured by applied researchers fails to forecast certain long run components.

Two basic features characterize UC decompositions. First, a researcher specifies flexible structures for the trend, the cycle and the other features of the data. These structures in turns imply an ARIMA representation for y_t which is more complicated than the one typically selected with standard methods. Second, given the assumed structure, the data is allowed to select the characteristics of the components and diagnostic testing can be employed to examine what is left unexplained.

For most of the discussion, we assume that there are only two unobservable components, i.e. $y_t = y_t^x + y_t^c$. Extensions to series containing, e.g., seasonals or irregular are immediate and left as an exercise to the reader. Assume that y_t^x can be represented as a random walk with drift:

$$y_t^x = \bar{y} + y_{t-1}^x + e_t^x \qquad e_t^x \sim iid\ (0, \sigma_x^2) \tag{3.4}$$

and that (y_t^c, e_t^x) is a jointly covariance stationary process. Note that if $\sigma_x^2 = 0\ \forall t$, y_t^x is a linear trend. While for the rest of this subsection we restrict attention to (3.4), more general trend specifications are possible. For example, "cyclical" trend movements are obtained by setting $y_t^x = \bar{y} + y_{t-1}^x + y_{t-1}^c + e_t^x$, $e_t^x \sim iid\ (0, \sigma_x^2)$. This specification makes trend and cycle correlated and was shown to fit US data better than (3.4) plus a simple cycle specification (see Harvey (1985)). Trends with higher order of integration are obtained if $\bar{y}$ drifts itself as a random walk.

Exercise 3.7 *(Harvey and Jeager) Suppose*

$$y_t^x = y_{t-1}^x + \bar{y}_{t-1} + e_t^x \qquad e_t^x \sim \ iid\ (0, \sigma_x^2) \tag{3.5}$$
$$\bar{y}_t = \bar{y}_{t-1} + v_t \qquad v_t \sim \ iid\ (0, \sigma_v^2) \tag{3.6}$$

Show that if $\sigma_v^2 > 0$ and $\sigma_x^2 = 0$, y_t^x is an I(2) process. Under what conditions is y_t "smooth", i.e. $\Delta^2 y_t$ are small? Verify that if $\sigma_v^2 = 0$, (3.5)-(3.6) collapse to (3.4) and that if $\sigma_v^2 = \sigma_x^2 = 0$, the trend is deterministic.

To complete the specification we need to postulate a process for y_t^c and the relationship between y_t^c and e_t^x. There are three possibilities. In the first case we assume

$$y_t^c = D^c(\ell) e_t^c \qquad e_t^c \sim \ iid\ \ (0, \sigma_c^2) \tag{3.7}$$

where $D^c(\ell) = 1 + D_1^c \ell + \ldots$ and e_t^c is orthogonal to $e_{t-\tau}^x$, $\forall \tau$. Then (3.1)-(3.4)-(3.7) imply:

$$D(\ell) e_t = e_t^x + (1-\ell) D(\ell)^c e_t^c \tag{3.8}$$

so that $|D(1)|^2 \sigma_e^2 = \sigma_x^2$ and the coefficients of $D(\ell)^c$ can then be found using $D(\ell)D(\ell^{-1})\sigma_e^2 - \sigma_x^2 = (1-\ell)(1-\ell^{-1}) D(\ell)^c D(\ell^{-1})^c \sigma_c^2$, provided that the roots of $D^c(\ell)$ are on or outside the complex unit circle. Note that, as in the BN decomposition, the spectral density of Δy_t^c has zero power at the zero frequency. Hence, (3.8) places restrictions on y_t as we show next.

Exercise 3.8 *Show that when the model is composed of (3.1)-(3.4)-(3.7) $\mathcal{S}_{\Delta y_t}(\omega)$ has a global minimum at $\omega = 0$. Conclude that y_t can not be represented as an ARIMA(1,1,0) with high autoregressive root.*

Since the restrictions imposed by (3.8) may not be appropriate for all y_t, we want to have other cyclical structures to describe models of the form (3.1). A second representation is

$$y_t^c = D^{cx}(\ell) e_t^x \tag{3.9}$$

where $D^{cx}(\ell) = 1 + D_1^{cx}\ell + \ldots$. In (3.9) innovations to the trend and to the cycle are perfectly correlated. Note that while the orthogonality of e_t^c and e_t^x restricts the ARIMA processes suitable to represent y_t, perfect correlation of the two innovations place no testable constraints on the ARIMA model for y_t. In particular, it is no longer true that an ARIMA(1,1,0) is an unlikely representation for y_t. A third representation for y_t^c is

$$y_t^c = D^c(\ell) e_t^c + D^{cx}(\ell) e_t^x \tag{3.10}$$

This specification is observationally equivalent to the cyclical-trend model discussed above.

While it is typical to specify an AR process for y_t^c, one could also choose trigonometric functions. Such representations are useful if one is interested in emphasizing a particular frequency where the cycle may have most of its power. For example, one could set

$$y_t^c = \frac{(1 - \rho_y \cos\omega\ell) e_t^{1c} + (\rho_y \sin\omega\ell) e_t^{2c}}{1 - 2\rho_y \cos\omega\ell + \rho_y^2 \ell^2} \tag{3.11}$$

where $e_t^{ic} \sim \ iid\ (0, \sigma_{e_i}^2), i = 1, 2$, $0 \leq \rho_y \leq 1$ and $0 \leq \omega \leq \pi$.

Exercise 3.9 *Show that y_t^c in (3.11) is an ARMA(2,1) process. Show that it reduces to an AR(2) if $\sigma_{e_1}^2 = 0$. Note that for $0 < \omega < \pi$, the roots of the AR(2) polynomial are complex with modulus ρ_y. Finally, show that for $\omega = 0$ or $\omega = \pi$, y_t^c is an AR(1) process.*

Since cycles at frequency ω_i specified via (3.11) are orthogonal to cycles at frequency $\omega_{i'}$ when ω_i and $\omega_{i'}$ are Fourier frequencies, cycles of multiple length can be accounted for by taking a linear combination of (3.11) at any two frequencies.

Example 3.6 *Suppose we are convinced that cycles in the data have changed average periodicity over time from, say, 8 to 6 years. If quarterly data are available then $y_t^c = y_t^{c_1} + y_t^{c_2} = \frac{(1-\rho_y \cos\omega_1 \ell)e_t^{1c} + (\rho_y \sin\omega_1 \ell)e_t^{2c}}{1-2\rho_y \cos\omega_1 \ell + \rho_y^2 \ell^2} + \frac{(1-\rho_y \cos\omega_2 \ell)e_t^{1c} + (\rho_y \sin\omega_2 \ell)e_t^{2c}}{1-2\rho_y \cos\omega_2 \ell + \rho_y^2 \ell^2}$ where $\omega_1 = \frac{2\pi}{32}$ and $\omega_2 = \frac{2\pi}{24}$.*

Given (3.4) and a model for y_t^c, it is immediate to show that y_t has an ARIMA format.

Example 3.7 *Consider the trend specification (3.4), the trigonometric cycle specification (3.11) and assume that $y_t = y_t^x + y_t^c + e_t$ where $e_t \sim$ iid $(0, \sigma_e^2)$. Then $\Delta y_t = \bar{y} + e_t^x + \Delta y_t^c + \Delta e_t$. Therefore if y_t^c is an ARMA(2,1), Δy_t is a restricted ARMA(2,3). The restrictions insure that i) a cycle if it exists can be found and ii) the local identifiability of the various components (if $\rho_y > 0$, there are no common factors in the AR and MA parts).*

Two methods are typically employed to obtain estimates of y_t^x in UC models: Linear Minimum Mean Square (LMMS) and the Kalman filter. The Kalman filter will be discussed in chapter 6. To obtain LMMS estimates, we let $\mathcal{F}_{-\infty}^{\infty} = \{\ldots, y_{-1}, y_0, y_1, \ldots,\}$ and use the Wiener-Kolmogorov prediction formulas (see e.g Whittle (1980). Then $y_t^x = \mathcal{B}^x(\ell) y_t$, where $\mathcal{B}^x(\ell)$ is two-sided and, for a model composed of (3.1)-(3.4)-(3.10), given by: $\mathcal{B}^x(\ell) = \sigma_x^2[1 + (1-\ell^{-1})D^{cx}(\ell^{-1})][D(\ell)D(\ell^{-1})\sigma_y^2]^{-1}$. Since only $\mathcal{F}_0^\tau = \{y_0, y_1, \ldots, y_\tau\}$ is available, define $y_t^x(\tau) \equiv E[y_t^x | \mathcal{F}_0^\tau]$. Then $y_t^x(\tau) = \sum_j \mathcal{B}_j^x E[y_{t-j} | \mathcal{F}_0^\tau]$ and estimates of the trend are obtained substituting unknown values of y_t with forecasts or backcasts constructed from $\mathcal{F}_0^\tau$. Clearly, $\mathcal{B}^x(\ell)$ depends on the model for y_t^c but differences across specifications arise only from the way future data is used to construct $y_t^x(\tau)$.

Exercise 3.10 *Show that estimates of $y_t^x(\tau)$ for all $\tau < t$ are the same regardless of whether (3.8),(3.9) or (3.10) is used.*

One implication of exercise 3.10 is that to obtain $y_t^c(t)$, it is sufficient to construct an ARIMA model for y_t, forecast in the distant future and set $y_t^c(t) = y_t - y_t^x(t)$, where $y_t^x(t)$ is the (forecast) estimate of the trend based on $\mathcal{F}_0^t$, adjusted for deterministic increases. Hence, $y_t^x(t)$ is similar to the permanent component obtained with the BN decomposition. However, as Morley, Nelson and Zivot (2002) have shown, this does not mean that the two cyclical components have similar (time series) properties.

Multivariate versions of UC decompositions have been initially suggested by Stock and Watson (1989)-(1990) and used by several other researchers. Multivariate UC decompositions typically impose the restriction that a y_t vector is driven in the long run by a reduced

number of permanent components; the transitory components, on the other hand, are allowed to be series specific. The multivariate UC setup is very close to the one employed in factor models (which we discuss in Chapter 11). In factor models there is an unobservable factor which captures the portion of the dynamics which are common to the series. Here the unobservable factor only captures the long run patterns in the data.

For a $m \times 1$ vector of integrated series, a multivariate UC decomposition is:

$$\begin{aligned} \Delta y_t &= \bar{y} + \mathbb{Q}(\ell)\Delta y_t^x + y_t^c && (3.12) \\ A^c(\ell) y_t^c &= e_t^c && (3.13) \\ A^x(\ell)\Delta y_t^x &= \bar{y}^x + e_t^x && (3.14) \end{aligned}$$

where y_t and y_t^c are $m \times 1$ vectors and y_t^x is a $m_1 \times 1$ vector, $m_1 < m$; while $A^c(\ell)$ and $A^x(\ell)$ are one-sided polynomial matrices in the lag operator.

There are two main identifying assumptions implicit in (3.12)-(3.14). First, the long run movements in y_t are driven by $m_1 < m$ processes. Second, $(y_{1t}^c, \ldots, y_{mt}^c, \Delta y_{1t}^x, \ldots, \Delta y_{m_1 t}^x)$ are uncorrelated at all leads and lags. Since it is impossible to separately identify $A^x(\ell)$ and $\mathbb{Q}(\ell)$ one typically sets $\mathbb{Q}(\ell) = \mathbb{Q}$ and assumes that at least one $\Delta y_{i't}^x$ enters each Δy_{it}. Note that when $A^x(\ell) \neq 1$, $y_t^x(t)$ is a $m_1 \times 1$ vector of coincident indicators, while $y_t^c(t)$ captures idiosyncratic movements.

Since the system (3.12)-(3.14) has a state space format, the unknown parameters ($A^c(\ell)$, $A^x(\ell), \bar{y}^x, \bar{y}, \mathbb{Q}, \Sigma_c, \Sigma_x$) and the unobservable components can be estimated by likelihood methods, recursively, with the Kalman filter. We defer the presentation of the Kalman filter recursions and of the prediction error decomposition of the likelihood to chapter 6.

3.1.4 Regime shifting decomposition

Although Hamilton's (1989) method has been devised to model recurrently segmented trends rather than to extract cycles, it naturally produces cyclical components which can be used as a benchmark to compare the properties of simulated DSGE models.

The idea of the approach is simple. Instead of choosing either a deterministic or a continuously changing stochastic specification, the trend is assumed to be regime specific, with the regime varying randomly over time. Within a regime, trend movements are deterministic. Two features of the approach need to be emphasized: (i) the resulting model for y_t is nonlinear in the conditional mean; (ii) shifts in the trend are driven by nonnormal errors.

For simplicity, we consider here only two regimes. Extensions to multiple regimes are straightforward and left as an exercise to the reader. Let Δy_t be stationary; let y_t^x and y_t^c be mutually independent and let $y_t^x = a_0 + a_1 \varkappa_t + y_{t-1}^x$, where $\varkappa_t \in (1, 0)$ is an unobservable two-state Markov chain indicator with $P[\varkappa_t = i | \varkappa_{t-1} = i'] = \begin{bmatrix} p_1 & 1-p_1 \\ 1-p_2 & p_2 \end{bmatrix}$, $p_1, p_2 < 1$. Because $\varkappa_t$ has a first order Markov structure we can rewrite it as

$$\varkappa_t = (1 - p_2) + (p_1 + p_2 - 1)\varkappa_{t-1} + e_t^x \qquad (3.15)$$

where e_t^x can take four values $[1-p_1, -p_1, -(1-p_2), p_2]$ with probabilities $[p_1, 1-p_1, p_2, 1-p_2]$.

Exercise 3.11 *Show that $\varkappa_t$ is covariance stationary. Is the process ergodic?*

The residuals e_t^x in (3.15) have two properties which we summarize next.

Exercise 3.12 *Show that (i) $E[e_t^x|\varkappa_{t-1}=i, i=0,1]=0$; (ii) $var[e_t^x|\varkappa_{t-1}=1]=p_1(1-p_1)$ and $var[e_t^x|\varkappa_{t-1}=0]=p_2(1-p_2)$.*

Exercise 3.12 shows that e_t^x are uncorrelated with previous realizations of the state but not independent. Note that if e_t^x were normal, uncorrelation implies independence. Therefore, the particular structure present in e_t^x implies separation between the two concepts.

The non-independence of e_t^x implies non-independence of $\varkappa_t$. Solving backward (3.15) and taking expectations at time zero we have $E_0\varkappa_t = \frac{1-p_2}{2-p_1-p_2}(1-(p_1+p_2-1)^t)+(p_1+p_2-1)^t E_0\varkappa_0$ and, taking limits, $\lim_{t\to\infty} E_0\varkappa_t = \frac{1-p_2}{2-p_1-p_2}\equiv\tilde{p}$. Let $P[\varkappa_t=1|\mathcal{F}_0]=p^0$.

Exercise 3.13 *Show that $var_0\varkappa_t = \frac{var[e_t^x|\varkappa_0]}{(2-p_1-p_2)^2}(1-(p_1+p_2-1)^t)^2+(p_1+p_2-1)^{2t}E[\varkappa_0-E\varkappa_0]^2$. Compute $\lim_{t\to\infty} var_0\varkappa_t$ and show that it is not statistically independent of $\varkappa_{t-j}$.*

Exercise 3.13 shows that the moment structure of $\varkappa_t$ (and therefore y_t^x) is nonlinear. This is important for forecasting. In fact, the model for y_t^x can be rewritten as:

$$(1-(p_1+p_2-1)\ell)\Delta y_t^x = a_1(1-(p_1+p_2-1)\ell)\varkappa_t = a_1(1-p_2)+a_0(2-p_1-p_2)+e_t^x \quad (3.16)$$

Hence, although y_t^x looks like an ARIMA(1,1,0) structure, forecasts of $y_{t+\tau}^x, \tau\geq 1$ based on a such a model are suboptimal since the non-linear structure in e_t^x is ignored. In fact, the optimal trend forecasts are:

$$E_t\Delta y_{t+\tau}^x = a_0 + a_1 E_t\varkappa_{t+\tau} = a_0 + a_1[\tilde{p}+(p_1+p_2-1)^t(P[\varkappa_t=1|\mathcal{F}_t]-\tilde{p})] \quad (3.17)$$

where $\mathcal{F}_t$ represents the information set. (3.17) is optimal since it incorporates the information that y_t^x changes only occasionally due to the discrete shifts in e_t^x.

Exercise 3.14 *Let $\bar{\varkappa}_t \equiv \sum_{j=1}^t \varkappa_j$, i.e. $\bar{\varkappa}_t$ is the cumulative number of ones. Show (i) $y_t^x = y_0^x + a_1\bar{\varkappa}_t + a_0 t$; (ii) $E_0[y_t^x|y_0^x,p^0] = y_0^x + a_1[\tilde{p}t + \sum_{j=1}^t (p_1+p_2-1)^j(p^0-\tilde{p})] + a_0 t$ and (iii) $\lim_{t\to\infty} E_0[y_t^x - y_{t-1}^x|y_0^x,p^0] = a_0 + a_1\tilde{p}$.*

Exercise 3.14 indicates that the growth rate of y_t^x is asymptotically independent of the information at time 0. Intuitively, as $t\to\infty$, y_t will be in the growth state a_0+a_1 with probability $\tilde{p}$ and in the growth state a_0 with probability $(1-\tilde{p})$. Note also that information concerning the initial state has permanent effects on y_t^x. In fact, $p^0-\tilde{p}\neq 0$ produces permanent changes in y_t^x.

To complete the specification, a process for y_t^c needs to be selected.

Example 3.8 *Suppose $y_t^c \sim$ iid $\mathbb{N}(0,\sigma_c^2)$. Then y_t has the representation $(1-(p_1+p_2-1)\ell)\Delta y_t = a_1(1-p_2)+a_0(2-p_1-p_2)+e_t-D_1e_{t-1}$ with $e_t-D_1e_{t-1} = e_t^x+y_t^c-(p_1+p_2-1)y_{t-1}^c$, where D_1 and σ_e^2 satisfy $(1-D_1^2)\sigma_e^2 = (1-(p_1+p_2-1)^2)\sigma_c^2+\sigma_x^2$ and $D_1^2\sigma_e^2 = (p_1+p_2-1)^2\sigma_c^2$.*

Exercise 3.15 *Suppose that* $\Delta y_t^c = A^c(\ell)\Delta y_{t-1}^c + e_t^c$, *where* $e_t^c \sim$ *iid* $\mathbb{N}(0, \sigma_c^2)$, $\forall \tau > 1$ *and* $A^c(\ell)$ *is of order* q_c. *Show the implied model for* y_t.

Given (3.15) and a model for y_t^c (for example the one of exercise 3.15), our task is to estimate the unknown parameters and to obtain an estimate of $\varkappa_t$. First, we consider a recursive algorithm to estimate $\varkappa_t$. That is, given $P[\varkappa_{t-1} = \bar{\varkappa}_{t-1}, \varkappa_{t-2} = \bar{\varkappa}_{t-2}, \ldots, \varkappa_{t-\tau} = \bar{\varkappa}_{t-\tau}|y_{t-1}, y_{t-2}, \ldots]$, we want $P[\varkappa = \bar{\varkappa}_t, \varkappa_{t-1} = \bar{\varkappa}_{t-1}, \ldots, \varkappa_{t-\tau+1} = \bar{\varkappa}_{t-\tau+1}|y_t, y_{t-1}, \ldots]$. The algorithm consists of 5 steps.

Algorithm 3.1

1) *Compute* $P[\varkappa = \bar{\varkappa}_t, \varkappa_{t-1} = \bar{\varkappa}_{t-1}, \ldots, \varkappa_{t-\tau} = \bar{\varkappa}_{t-\tau}|\Delta y_{t-1}, \Delta y_{t-2}, \ldots] = P[\varkappa = \bar{\varkappa}_t|\varkappa_{t-1} = \bar{\varkappa}_{t-1}]P[\varkappa_{t-1} = \bar{\varkappa}_{t-1}, \varkappa_{t-2} = \bar{\varkappa}_{t-2}, \ldots, \varkappa_{t-\tau} = \bar{\varkappa}_{t-\tau}|\Delta y_{t-1}, \Delta y_{t-2}, \ldots]$ *where* $P[\varkappa_t = \bar{\varkappa}_t|\varkappa_{t-1} = \bar{\varkappa}_{t-1}]$ *is the transition matrix of* $\varkappa_t$.

2) *Compute the joint probability of* Δy_t *and* $\{\varkappa_j\}_{j=t-\tau}^t$, *i.e.:* $f(\Delta y_t, \varkappa_t = \bar{\varkappa}_t, \varkappa_{t-1} = \bar{\varkappa}_{t-1}, \ldots, \varkappa_{t-\tau} = \bar{\varkappa}_{t-\tau}|\Delta y_{t-1}, \Delta y_{t-2}, \ldots) = f(\Delta y_t|\varkappa = \bar{\varkappa}_t, \varkappa_{t-1} = \bar{\varkappa}_{t-1}, \ldots, \Delta y_{t-1}, \Delta y_{t-2}, \ldots)P[\varkappa = \bar{\varkappa}_t, \varkappa_{t-1} = \bar{\varkappa}_{t-1}, \ldots, \varkappa_{t-\tau} = \bar{\varkappa}_{t-\tau}|\Delta y_{t-1}, \Delta y_{t-2}, \ldots]$ *where* $f(\Delta y_t|\varkappa_t = \bar{\varkappa}_t, \varkappa_{t-1} = \bar{\varkappa}_{t-1}, \ldots, \Delta y_{t-1}, \Delta y_{t-2}, \ldots) = \frac{1}{\sqrt{2\pi}\sigma_c} exp[-\frac{1}{2\sigma_c^2}((\Delta y_t - a_0 - a_1\varkappa_t)(1 - A^c(\ell))^2]$.

3) *Compute* $f(\Delta y_t|\Delta y_{t-1}, \Delta y_{t-2}, \ldots) = \sum_{\varkappa_t=0}^1 \sum_{\varkappa_{t-1}=0}^1 \cdots \sum_{\varkappa_{t-\tau}=0}^1 f(\Delta y_t, \varkappa = \bar{\varkappa}_t, \varkappa_{t-1} = \bar{\varkappa}_{t-1}, \ldots, \varkappa_{t-\tau} = \bar{\varkappa}_{t-\tau}|\Delta y_{t-1}, \Delta y_{t-2}, \ldots)$. *This is the predictive density of* Δy_t *based on* $t-1$ *information.*

4) *Apply Bayes theorem to obtain:* $P(\varkappa_t = \bar{\varkappa}_t, \varkappa_{t-1} = \bar{\varkappa}_{t-1}, \ldots|\Delta y_t, \Delta y_{t-1}, \Delta y_{t-2}, \ldots) = \frac{f(\Delta y_t, \varkappa_t = \bar{\varkappa}_t, \varkappa_{t-1} = \bar{\varkappa}_{t-1}, \ldots, \varkappa_{t-\tau} = \bar{\varkappa}_{t-\tau}|\Delta y_{t-1}, \Delta y_{t-2}, \ldots)}{f(\Delta y_t|\Delta y_{t-1}, \Delta y_{t-2}, \ldots)}$.

5) *Obtain* $P[\varkappa_t = \bar{\varkappa}_t, \varkappa_{t-1} = \bar{\varkappa}_{t-1}, \ldots, \varkappa_{t-\tau+1} = \bar{\varkappa}_{t-\tau+1}|\Delta y_t, \Delta y_{t-1}, \Delta y_{t-2}, \ldots] = \sum_{\varkappa_{\tau-r}=0}^1 P[\varkappa_t = \bar{\varkappa}_t, \varkappa_{t-1} = \bar{\varkappa}_{t-1}, \ldots, \varkappa_{t-\tau} = \bar{\varkappa}_{t-\tau}|\Delta y_t, \Delta y_{t-1}, \Delta y_{t-2}, \ldots]$.

To start the algorithm one needs $P[\varkappa_0 = \bar{\varkappa}_0, \varkappa_{-1} = \bar{\varkappa}_{-1}, \ldots, \varkappa_{-\tau+1} = \bar{\varkappa}_{-\tau+1}|y_0, y_{-1}, \ldots]$. If this is unknown, one can use $P[\varkappa_0 = \bar{\varkappa}_0, \varkappa_{-1} = \bar{\varkappa}_{-1}, \ldots, \varkappa_{-\tau+1} = \bar{\varkappa}_{-\tau+1}]$, the unconditional probability of the $\tau - 1$ histories of $\varkappa_0$, which is obtained setting $P[\varkappa_{-\tau+1} = 1] = \tilde{p}$, $P[\varkappa_{-\tau+1} = 0] = 1 - \tilde{p}$ and recursively constructing $P[\varkappa_{-\tau+j}]$ $j = 2, 3, \ldots$ using step 1) of the algorithm. Alternatively, one could treat $P[\varkappa_{-\tau+1} = 1]$ as a parameter to be estimated.

Extensions of the basic setup are considered in the next exercise

Exercise 3.16 *i) Suppose there are* n *states so that the input of the algorithm consists of* n^τ *elements. Write down the algorithm to estimate* $\varkappa_t$ *with information up to* $t-1$.
ii) Let $A^c(\ell) = A^c(\ell, \varkappa_t)$. *Write down the algorithm to estimate* $\varkappa_t$.
iii) Let $\sigma_c^2 = \sigma_c^2(\varkappa_t)$. *Write down the algorithm to estimate* $\varkappa_t$.

Example 3.9 *One interesting extension is obtained when the probability of switching states depends on observable variables. For example, set* $P(\varkappa_t = i|\varkappa_{t-1} = i, x_{t-1}\alpha_i) = \frac{exp(x'_{t-1}\alpha_i)}{1+exp(x'_{t-1}\alpha_i)}$ *and* $P(\varkappa = i|\varkappa_{t-1} = i', x_{t-1}\alpha_{i'}) = 1 - P(\varkappa = i|\varkappa_{t-1} = i, x_{t-1}\alpha_i)$*, where* $x_{t-1} = (1, y_{1,t-1}, \ldots, y_{q,t-1}), \alpha_i = (\alpha_{i0}, \alpha_{i1}, \ldots, \alpha_{iq})$ *and let* $f(\Delta y_t|\varkappa, \theta) = \frac{1}{\sqrt{2\pi}\sigma_c} exp[\frac{-(\Delta y_t(\varkappa) - \overline{\Delta y}(\varkappa))^2}{2\sigma_c^2}]$*. Then* Δy_t *has potentially a switching mean and the probability of switching is time dependent. Hence there may be duration dependence in the fluctuations of* Δy_t*.*

Next, we consider parameters estimation. Since in step 3) of algorithm 3.1 we constructed the likelihood of the $t - th$ observation, parameters estimate can be obtained once algorithm 3.1 has been run for all t's, summing the log of these likelihoods. In fact, $\ln f(\Delta y_t, \Delta y_{t-1}, \ldots,$
$\Delta y_1|\Delta y_0, \ldots, \Delta y_{-\tau+1}) = \sum_{t=1}^{T} \ln f(\Delta y_t|\Delta y_{t-1}, \ldots, \Delta y_1, \ldots, \Delta y_{-\tau+1})$ can be numerically maximized with respect to $(\alpha_0, \alpha_1, p_1, p_2, A_j)$ (and $P[\varkappa_{-\tau+1} = 1]$ if needed).

From step 4) of the algorithm, we can also infer $\varkappa_t$ given current and past values of y_t, integrating out $\varkappa_{t-j}, j \geq 1$, i.e., $P[\varkappa_t = \bar{\varkappa}_t|\Delta y_t, \Delta y_{t-1}, \ldots] = \sum_{\varkappa_{t-1}=0}^{1}, \ldots, \sum_{\varkappa_{t-\tau}=0}^{1} P[\varkappa = \bar{\varkappa}_t, \varkappa_{t-1} = \bar{\varkappa}_{t-1}, \ldots|\Delta y_t, \Delta y_{t-1}, \ldots]$. This could be useful, e.g., to decide whether at some date the economy was in a recession or not.

Example 3.10 *Step 4) of algorithm 3.1 can also be used to evaluate the ex-post probability that* $\varkappa_{t-j} = \bar{\varkappa}_{t-j}$ *(given time t information). For example, we could calculate the probability that in 1975:1 we were in the low growth state, given information up to, say, 2003:4. This involves, computing* $P[\varkappa_{\bar{t}} = \bar{\varkappa}_{\bar{t}}|\Delta y_t, \Delta y_{t-1}, \ldots]$ *for some* $t - \tau \leq \bar{t} \leq t$*.*

The above framework is easy to manipulate but it unrealistically assumes that the cyclical component has a unit root. Eliminating this unit root brings realism to the specification but substantially complicates the calculations since the algorithm has to keep track of the entire past history of $\varkappa_t$. To illustrate the point suppose $y_t = y_t^x + y_t^c$ and

$$y_t^x = y_{t-1}^x + a_0 + a_1\varkappa_t \tag{3.18}$$

$$(1 - A^c(\ell))y_t^c = e_t^c, \quad e_t^c \sim \ iid\ \mathbb{N}(0, \sigma_c^2) \tag{3.19}$$

where $\varkappa_t$ is a two-state Markov chain. Then $\Delta y_t = a_0 + a_1\varkappa_t + \Delta y_t^c$ and solving backward $y_t^c = (\sum_{i=1}^{t} \Delta y_i - a_0 t - a_1 \sum_{i=1}^{t} \varkappa_i) + y_0^c$ and $e_t^c = (1 - A_1^c\ell - A_2^c\ell^2, \ldots, -A_{q_c}^c\ell^{q_c})[\sum_{i=1}^{t} \Delta y_i - a_0 t] + (1 - A_1^c - A_2^c, \ldots, -A_{q_c}^c)y_0^c - a_1(1 - A_1^c - A_2^c - \ldots - A_{q_c}^c)\sum_i \varkappa_i + a_1 \sum_{j=1}^{q_c}(\sum_{i=j}^{q_c} A_i^c)\varkappa_{t-j+1}$.

If $A^c(\ell)$ has a unit root $(1 - A^c(1)) = 0$ so that $e_t^c = (1 - A_1^c\ell - A_2^c\ell^2 - \ldots)[\sum_i \Delta y_i - a_0 t] + a_1 \sum_j (\sum_i A_i^c)\varkappa_{t-j+1}$ which is the same expression obtained when y_t^c is an ARIMA(q_c,1,0). If it is not the case, the entire history of $\varkappa_t$ becomes a state variable for the problem. To solve this computation problem, Lam (1990) uses the sum of $\varkappa_t$ as a state variable. Note that since y_0^c affects the likelihood, it is treated as a parameter to be estimated.

Exercise 3.17 *Modify algorithm 3.1 to allow* $(\sum_{i=1}^{t} \varkappa_i)$ *to be a new state variable.*

Note that the probability distribution of $\sum_{i=1}^{t} \varkappa_i$ can be computed as $P[\sum_i \varkappa_i = \bar{\bar{\varkappa}}_i|\Delta y_t, \ldots, \Delta y_1] = \sum_{\varkappa_t=0}^{1} \ldots \sum_{\varkappa_{t-\tau}=0}^{1} P[\varkappa_t = \bar{\varkappa}_t, \ldots, \varkappa_{t-\tau} = \bar{\varkappa}_{t-\tau} \sum_i \varkappa_i = \bar{\bar{\varkappa}}_i|\Delta y_t, \ldots, \Delta y_1]$. An estimate of y_t^c is $\hat{y}_t^c = \sum_{i=1}^{t} \Delta y_i - a_0 t + y_0^c - a_1 \sum_{j=0}^{t} \bar{\bar{\varkappa}} P[\sum_i \varkappa_i = \bar{\bar{\varkappa}}|\Delta y_t, \Delta y_{t-1}, \ldots, \Delta y_1]$ and, given y_0^c, an estimate of the Markov trend is $\hat{y}_t^x = y_t - \hat{y}_t^c$.

3.2 Hybrid Decompositions

3.2.1 The Hodrick and Prescott (HP) Filter

The HP filter has been and still is one of the preferred methods to extract cyclical components from economic time series. Two basic features characterize HP decompositions. First, trend and cycle are assumed to be uncorrelated. Second, the trend is assumed to be a "smooth" process, that is, it is allowed to change over time as long as the changes are not abrupt. Hodrick and Prescott make the "smoothness' concept operational by penalizing variations in the second difference of the trend. Under these conditions y_t^x can be identified and estimated using the following program

$$\min_{y_t^x} \{\sum_{t=0}^{x} (y_t - y_t^x)^2 + \lambda \sum_{t=0}^{T} ((y_{t+1}^x - y_t^x) - (y_t^x - y_{t-1}^x))^2\} \tag{3.20}$$

where λ is a parameter controlling the smoothness of the trend. As λ increases y_t^x becomes smoother and for $\lambda \to \infty$, it becomes linear. A program like (3.20) can be formally derived as follows. Let the cyclical component and the second difference of the trend be white noises. Then weighted least square minimization leads to $\min_{\{y_t^x\}} \{\sum_{t=0}^{T} \frac{(y_t - y_t^x)^2}{\sigma_{y^c}^2} + \frac{\sum_{t=0}^{T} (y_{t+1}^x - 2y_t^x + y_{t-1}^x)^2}{\sigma_{\Delta^2 y^x}^2}\}$, which produces $\lambda = \frac{\sigma_{y^c}^2}{\sigma_{\Delta^2 y^x}^2}$, where $\sigma_{\Delta^2 y^x}^2$ is the variance of the innovations in the second difference of the trend and $\sigma_{y^c}^2$ is the variance of the innovations in the cycle. When $\lambda = \frac{\sigma_{y^c}^2}{\sigma_{\Delta^2 y^x}^2}$, Wabha (1980) shows that (3.20) defines the best curve in a cloud of points, in the sense of making the mean square of the fitting error as small as possible.

Interestingly, the trend produced by (3.20) is identical to the one produced by a UC decomposition where the drift in the trend is itself a random walk and the cyclical component is a white noise (see Harvey and Jeager (1993)) . In particular, if we set $\lambda = \frac{\sigma_c^2}{\sigma_v^2}$, restrict $\sigma_x^2 = 0$ in the setup of exercise 3.7 and let $T \to \infty$, (3.20) is the optimal signal extraction method to recover y_t^x (see e.g. Gomez (1999)). Clearly, if y_t is not simply trend plus noise or when the cyclical component is not iid, the filter is no longer optimal. Hence, rather than estimating λ, the literature selects it a-priori so as to carve out particular frequencies of the spectrum. For example, the value $\lambda = 1600$ typically used for quarterly data, implies that the standard error of the cycle is 40 times larger than the standard error of the second difference of the trend and this, in turns, implies that cycles longer than 6-7 years are attributed to the trend. The choice of λ is not necessarily innocuous and implicit estimates of λ obtained using BN or UC decompositions are only in the range [2, 8].

Exercise 3.18 *Show that the solution to (3.20) is* $y^x = (\mathbb{F}^{HP})^{-1}y$, *where* $y = [y_1, y_2, \ldots, y_T]'$, $y_t^x = [y_1^x, y_2^x, \ldots, y_T^x]'$. *Display the* $T \times T$ *matrix* $\mathbb{F}^{HP}$.

For quarterly data $\mathbb{F}^{HP}$ has a particular form: only $t-2, t-1, t, t+1, t+2$ observations at each t matter in constructing y_t^x and the weights on leads and lags of y_t depend on λ but are symmetric. Therefore, the HP trend extractor is a two-sided, symmetric moving average filter. Once y_t^x is available, an estimate of the cyclical component is $y_t^c = y_t - y_t^x$.

The filter defined by exercise 3.18 is time dependent. Furthermore, its two-sided nature creates beginning and end-of-the-sample problems. In fact, the elements of $\mathbb{F}^{HP}$ for the initial two and the final two observations differ from those of observations in the middle of the sample. This creates distortions when one is interested in the properties of y_t around the end of the sample. One unsatisfactory solution is to throw away these observations when constructing interesting statistics. Alternatively, one could dump the effects of these observations by appropriately weighting them in the computation of the autocovariance function of the filtered data. The preferred solution is to use a version of (3.20) where t runs from $-\infty$ to $+\infty$. This modified problem defines a set of linear, time invariant weights which, away from the beginning and the end, are close to those obtained in exercise 3.18.

The modified minimization problem produces an estimate of the cycle of the form $y_t^c = \mathcal{B}(\ell)(1-\ell)^4 y_t \equiv \mathcal{B}^c(\ell)y_t$, where $\mathcal{B}(\ell)$ is (see e.g. Cogley and Nason (1995)):

$$\mathcal{B}(\ell) = \frac{|\lambda_1|^2}{\ell^2}[1 - 2Re(\lambda_1)\ell + |\lambda_1|^2\ell^2]^{-1}[1 - 2Re(\lambda_1)\ell^{-1} + |\lambda_1|^2\ell^{-2}]^{-1} \tag{3.21}$$

λ_1^{-1} is the stable root of $[\lambda^{-1}\ell^2 + (1-\ell)^4]$, $Re(\lambda_1)$ is the real part of λ_1 and $|\lambda_1|^2$ is its squared modulus. Note that when $\lambda = 1600, Re(\lambda_1) = 0.89$, $|\lambda_1|^2 \simeq 0.8$ and the cyclical weights $\mathcal{B}_j^c, j = -\infty, \ldots, 0, \ldots, \infty$ can be written as (see Miller (1976)):

$$\begin{aligned} \mathcal{B}_j^c &= -(0.8941)^j[0.0561 * cos(0.1116j) + 0.0558 * sin(0.1116j)] \quad j \neq 0 & (3.22)\\ &= 1 - [0.0561 * cos(0) + 0.0558 * sin(0)] \quad j = 0 & (3.23) \end{aligned}$$

The cyclical filter can also be written as (see King and Rebelo (1993)):

$$\mathcal{B}^c(\ell) = \frac{(1-\ell)^2(1-\ell^{-1})^2}{\frac{1}{\lambda} + (1-\ell)^2(1-\ell^{-1})^2} \tag{3.24}$$

Figure 3.1 plots $\mathcal{B}_j^c$ from (3.22)-(3.23) and its gain function for $\lambda = 100, 400, 1600, 6400$. The weights have a sharp bell shape appearance, with the first time crossing of the zero line at lag 2 and again around lag 20.

For stationary y_t, increasing λ, adds to y_t^c cycles with longer and longer periodicity. Alternatively, since the area under the spectrum is the variance of y_t, increasing λ increases the importance of y_t^c relative to y_t^x.

To study the properties of the cyclical HP filter, it is worth distinguishing whether y_t is covariance stationary or integrated (and of what order). When y_t is stationary and $\lambda = 1600$, the gain of $\mathcal{B}^c(\ell)$ is $Ga^0(\omega) \simeq \frac{16\sin^4(\frac{\omega}{2})}{\frac{1}{1600}+16\sin^4(\frac{\omega}{2})} = \frac{4(1-\cos(\omega))^2}{\frac{1}{1600}+4(1-\cos(\omega))^2}$, which has the form depicted

in the top left panel of figure 3.2. It is immediate to notice that $Ga^0(\omega = 0) = 0$, so that the power of y_t at frequency zero goes to the trend. Furthermore, $Ga^0(\omega) \to 1$ for $\omega \to \pi$. Hence, the cyclical HP filter operates like a high pass filter, damping fluctuations with mean periodicity greater than 24 quarters per cycle ($\omega = 0.26$) and passing short cycles without changes. Because of this last feature, the cyclical HP filter leaves "undesirable" high frequency variability in y_t^c.

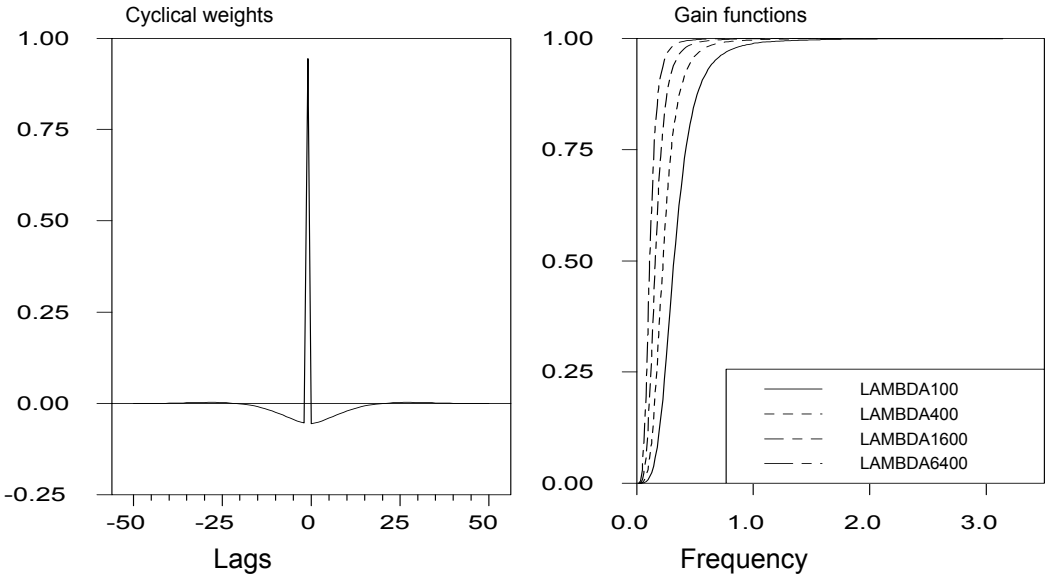


Figure 3.1: Cyclical weights and gain function, HP filter

When y_t is integrated, the cyclical HP filter has different properties. In fact, one can think of $\mathcal{B}^c(\ell)$ as a two-step filter: in the first step it renders y_t stationary; in the second it smooths the resulting stationary series with asymmetric moving average weights.

Example 3.11 *When y_t is integrated of order one, the first step filter is $(1-\ell)$ and the second is $\mathcal{B}(\ell)(1-\ell)^3$. When $\lambda = 1600$, the gain function of the latter is $Ga^1(\omega) \simeq [2(1-\cos(\omega))]^{-1} Ga^0(\omega)$ which has a peak at $\omega^* = \arccos[1 - \sqrt{\frac{0.75}{1600}}] = 0.21$ (roughly 7.6 years cycles) (see top central panel of figure 3.2). Hence, $Ga^1(\omega = 0) = 0$, but when applied to quarterly data, $\mathcal{B}(\ell)(1-\ell)^3$ damps long and short run growth cycles and strongly amplifies growth cycles at business cycle frequencies. For example, the variance of the cycles with average duration of 7.6 years is multiplied by 13 and the variance for cycles with periodicity between 3.2 and 13 years by a factor of 4.*

Exercise 3.19 *Suppose that y_t is I(2). What is the gain of $\mathcal{B}(\ell)(1-\ell)^2$? (call it $Ga^2(\omega)$).*

A plot of $Ga^2(\omega)$ when y_t is I(2) is in the top right panel of figure 3.2. Here $Ga^2(\omega = 0) = 0$ but the cyclical peak is very large. In fact, the variability of cycles corresponding to 60 periods or more is increased by 400 times. To summarize, the cyclical HP filter may induce spurious periodicity in integrated series. In particular, it may produce "periodic" cycles in series which have no power at business cycle frequencies (Yule-Slutsky effect).

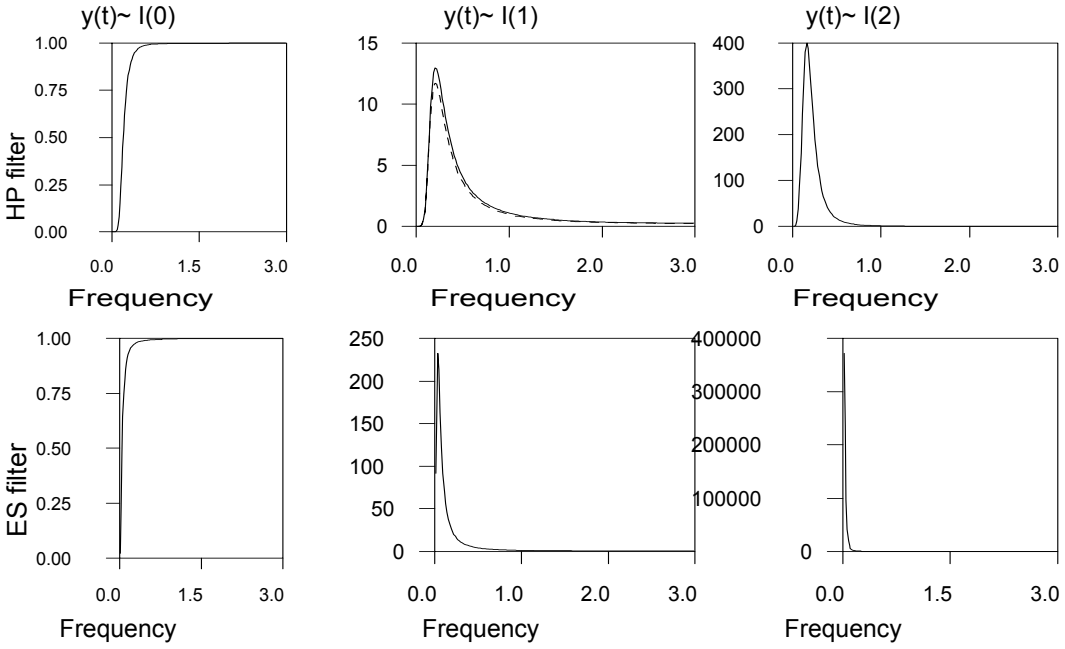


Figure 3.2: Gain functions, HP and ES filters

Example 3.12 *While macroeconomists agree that aggregate time series display persistent fluctuations, it is an open question whether they are integrated or not. Therefore, one may be tempted to dismiss the above arguments suggesting that a largest root around 0.95 is more probable than a root of 1.0. Unfortunately, for roots of the order of 0.95, the problem still remains. In fact, the cyclical HP filter is* $\mathcal{B}(\ell)(1-\ell)^3\frac{1-\ell}{1-0.95\ell}$. *The top middle panel of figure 3.2 plots the gain of this filter (dotted line in upper central panel). It is easy to see that the shape of the gain function and the magnitude of the amplification produced at business cycle frequencies is similar to the I(1) case.*

Exercise 3.20 *Suppose that* $y_t = a_0 + a_1 t + a_2 t^2 + e_t$. *Show that the HP filter eliminates linear and quadratic trends. (Hint: you can do this i) analytically; ii) applying the HP filter*

to a simulated process and explaining what is going on; iii) figuring out the spectral power of deterministic trends).

Exercise 3.21 *Consider the process* $y_t = 10 + 0.4t + e_t$ *where* $e_t = 0.8e_{t-1} + v_t$ *and* $v_t \sim$ *iid* $(0, 1)$. *Generate* y_t, $t = 1, \ldots, 200$, *and filter it with the HP filter. Repeat the exercise using* $y_t = \rho_y y_{t-1} + 10 + e_t$, *where* $\rho_y = 0.8, 0.9, 1.0$ *and* $y_0 = 10$. *Compare the autocovariance functions of* y_t^c *in the two cases. Is there any pattern in the results? Why?*

Because $\mathcal{B}^c(\ell)$ contains a term of the form $(1 - \ell)^4$, y_t^c will have, in general, a non-invertible MA representation (see chapter 4 for a definition of invertibility). For example, if y_t is stationary, y_t^c has four MA unit roots while if y_t is I(1), y_t^c has three MA unit roots. Non-invertibility implies that no finite AR representation for y_t^c exists. In other words, y_t^c will display strong serial correlation, regardless of whether y_t is serially correlated or not.

Example 3.13 *We have simulated data using* $y_t = 10 + 0.4t + e_t$, *where* $e_t = \rho_e e_{t-1} + v_t$, $v_t \sim$ *iid* $(0, 1)$ *and* $\rho_e = 0.4, 0.7, 1.0$. *Figure 3.3 reports the ACF functions of the true and the HP filtered cyclical component. Clearly, the higher is* ρ_e *the stronger is the persistence in the* y_t^c *and the longer it takes for the ACF to settle down at zero.*

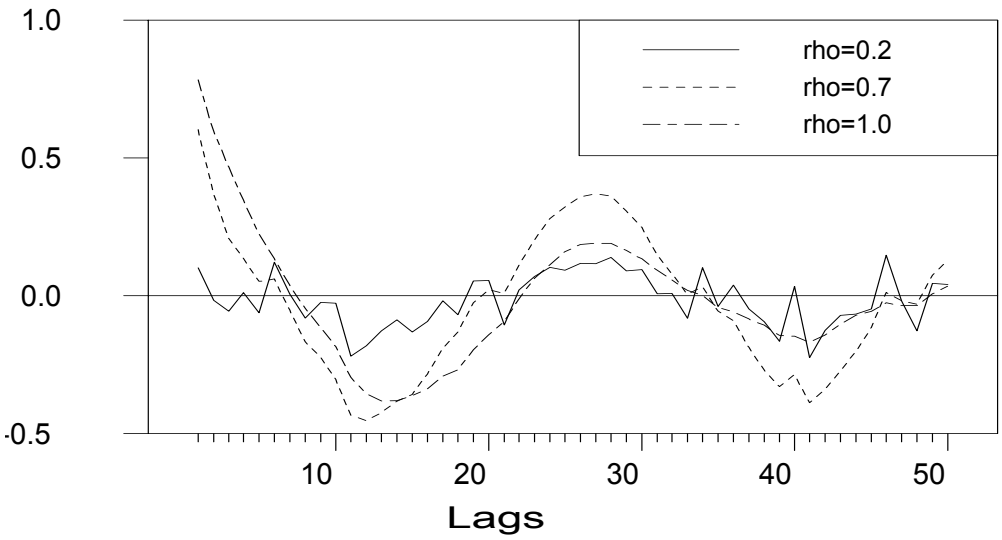


Figure 3.3: ACF of the cyclical component

The cyclical HP filter may not only induce artificial persistence or spurious periodicity; it may also create comovements that look like business cycle fluctuations in series which have no cycle. We show one extreme version of this phenomenon in the next exercise.

Exercise 3.22 *Let $y_{1t} = y_{1t-1} + e_{1t}$ and $y_{2t} = y_{2t-1} + e_{2t}$ and let $[e_{1t}, e_{2t}]' \sim (0, \Sigma_e)$.*
(i) Show that the spectral density matrix of $[\Delta y_{1t}, \Delta y_{2t}]$ is $\mathcal{S}(\omega) \propto \Sigma_e$.
(ii) Show that the spectral density matrix of $[y^c_{1t}, y^c_{2t}]'$ is $\mathcal{S}(\omega)Ga^1(\omega)$, where $Ga^1(\omega)$ is the gain of $\mathcal{B}(\ell)(1-\ell)^3$.
(iii) Let $\sigma_1 = \sigma_2 = 1$ and let $\sigma_{12} = 0.9, 0.5, 0.0$. Simulate y_{1t}, y_{2t} and plot $\mathcal{S}_{y^c}(\omega)$ in the three cases. Argue that when $\sigma_{12} \neq 0$, y^c_{1t} and y^c_{2t} display cycles of roughly the same periodicity as NBER cycles and that y^c_{1t} and y^c_{2t} have strong comovements.

Exercise 3.23 *Using quarterly US data for consumption and output, plot the spectral density matrix of $[\Delta c_t, \Delta GDP_t]'$. Plot the spectral density matrix of the HP detrended version of consumption and output. Describe the features of the plots and contrast them.*

In general, the use of the HP filter should be carefully monitored: uncritical use may produce a misleading impression of the ability of a model to reproduce the data. In particular, models which have little propagation mechanism and minor fluctuations, may acquire strong propagation and significant cyclical components once filtered with the HP filter (see Soderlin (1994) and Cogley and Nason (1995), for examples). Furthermore, as exercise 3.22 shows, even though the model and the data are only contemporaneous linked, application of the HP filter may make the similarities strong precisely at business cycle frequencies.

Example 3.14 *We have simulated 200 data points from the basic RBC model of chapter 2 with utility $U(c_t, c_{t-1}, N_t) = \frac{c_t^{1-\varphi}}{1-\varphi} + \ln(1 - N_t)$ assuming $\beta = 0.99, \varphi_c = 2.0, \delta = 0.025, \eta = 0.64$, steady state hours equal to 0.3. We log linearize the model and assume an AR(1) parameter equal to 0.9 and a variance equal to 0.0066 for the logarithm of the technology shock and an AR(1) parameter equal to 0.8 and a variance equal to 0.0146 for the goverment expenditure shock. Table 3.1 reports the mean of the cross correlation function of GDP with capital (K), real wage (W) and labor productivity (np) at lag zero and one and the mean standard deviations of the latter three variables calculated over 100 simulations, before and after HP filtering. Since the data generated by the model is stationary, the filtered statistics should be interpreted as describing the medium-high frequency properties of the simulated data. Clearly, both the relative ranking of variabilities and the size of the cross correlations are significantly different.*

	Raw data			*HP filtered data*		
Statistic	K_t	W_t	np_t	K_t	W_t	np_t
Correlation of GDP_t with	*0.49*	*0.65*	*0.09*	*0.84*	*0.95*	*-0.20*
Correlation of GDP_{t-1} with	*0.43*	*0.57*	*0.05*	*0.60*	*0.67*	*-0.38*
Standard deviation	*1.00*	*1.25*	*1.12*	*1.50*	*0.87*	*0.50*

Table 3.1: Simulated statistics

Since the smoothing parameter is chosen a-priori, one may wonder how to translate the value $\lambda = 1600$ into a value for monthly or annual data. For example, researchers have used

$\lambda = 400, 100, 10$ to compute y_t^c in annual data. Ravn and Uhlig (2002) show that insistence on the requirement that cycles of the same periodicity should be extracted, regardless of the frequency of the data, leads to select $\lambda = 129600$ for monthly data and $\lambda = 6.25$ for annual data when end-of-the-period data is used. While it is possible to derive these values analytically, we illustrate the logic for these choices by means of an example.

Example 3.15 *We generated 12000 (monthly) data points from an AR(1) process with* $\rho = 0.98$ *and variance of innovations equal to 1.0 and sampled it at quarterly and annual frequencies (using either end-of-the-period values or time averages) for a total of 4000 and 1000 observations. We have then applied the HP filter to monthly, quarterly and annual data where for quarterly data* $\lambda = 1600$*, for monthly data* $\lambda_j = 3^j\lambda$ *and for annual data* $\lambda_j = 0.25^j\lambda, j = 3, 4, 5$*. The variability of the quarterly HP filtered cycles are 2.20 (end-of-the period sampling) and 2.09 (averaging monthly data). The variability of the monthly series are 1.95 when j=3, 2.21 j=4, 2.48 for j=5. The variability of the annual series are 2.42 when j=3, 2.09 for j=4 and 1.64 for j=5 (end-of-the-period sampling) and 2.15 for j=3, 1.74 for j=4, and 1.33 for j=5 (time averaged data). Hence with end-of the period data,* $j = 4$ *is the most appropriate. For averaged data, j=4 or j=5 should probably be used.*

Exercise 3.24 *Let* $\mathcal{B}^c(\omega, \lambda)$ *be the cyclical HP filter for the quarterly data and let* $\mathcal{B}^c(\frac{\omega}{\tau}, \lambda_\tau)$ *the cyclical HP filter for the sampling frequency* $\frac{\omega}{\tau}$*, where* τ *measures the frequency of the observations relative to quarterly data, i.e.* $\tau = 0.25$ *for annual data and* $\tau = 3$ *for monthly data. Let* $\lambda_\tau = \tau^j\lambda$*. Calculate the gain function for monthly and annual data when* $j = 3.8, 3.9, 4, 4.1, 4.2$*. For which value of* j *is the gain function closer to the one for quarterly data?*

As we have seen, the HP filter is a mechanical device which defines the cycles it extracts via the selection of λ. In cross country comparisons the use of a single λ may be problematic since the mean length of domestic cycles in not necessarily the same. For example, if a country has cycles with average length of 9 years, mechanical application of the HP filter will move these cycles to the trend. The fact that quarterly HP filtered GDP data for Japan, Italy or Spain display very improbable expansions around the time of the first oil shock, when $\lambda = 1600$ is used, have prompted researchers to look for alternative ways to introduce smoothness in the trend. Marcet and Ravn (2000) suggested that for cross country comparisons one could either fix the amount of variability assigned to the trend or restrict the relative variability of the trend to the cycle. Roughly speaking, this amounts to making λ endogenous (as opposed to exogenously) when splitting the spectrum of y_t into components. The problem (3.20) in the latter case can be written as

$$\min_{y_t^x} \quad \sum_{t=1}^{T}(y_t - y_t^x)^2 \tag{3.25}$$

$$\mathcal{V}_1 \geq \frac{\sum_{t=1}^{T-2}[(y_{t+1}^x - y_t^x) - (y_t^x - y_{t-1}^x)]^2}{\sum_{t=1}^{T}(y_t - y_t^x)^2} \tag{3.26}$$

where $\mathcal{V}_1 \geq 0$ is a constant to be determined by the researcher, which measures the variability of the acceleration in the trend relative to the variability of the cyclical component. (3.25)-(3.26) and (3.20) are equivalent as the next exercise shows.

Exercise 3.25 *i) Show that if* $\mathcal{V}_1 = 0$, y_t^x *is a linear trend and if* $\mathcal{V}_1 \to \infty$, $y_t^x = y_t$.
ii) Let $\bar{\lambda}$ *be the (exogenous) value of* λ. *Show that the Lagrangian multiplier on (3.26) is* $\bar{\lambda} = \frac{\lambda}{(1-\lambda\mathcal{V}_1)}$. *Compute* λ *when* $\bar{\lambda} = 1600$ *and the ratio of variabilities is* $1, \frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{16}$.
iii) Show that a solution for $\mathcal{V}_1$ *can be found iterating on* $\mathcal{V}_1(\lambda) = \frac{\sum_{t=2}^{T-1}(y_{t+1}^x(\lambda)-2y_t^x(\lambda)+y_{t-1}^x(\lambda))^2}{\sum_{t=2}^{T-1}(y_t-y_t^x(\lambda))^2}$.

Intuitively, part ii) of exercise 3.25 indicates that if we want to make useful international comparisons (say, using the US as a benchmark) we should choose a λ that satisfy $\bar{\lambda} = \lambda(1-\mathcal{V}_1\lambda)^{-1}$, where $\bar{\lambda} = 1600$ and $\mathcal{V}_1$ is the relative variability of the two components in the US. Keeping $\mathcal{V}_1$ fixed across countries is more appealing than fixing λ since $\mathcal{V}_1$ is a parameter with some economic interpretation. Since in international comparisons it may not be clear which benchmark country one should use, one could substitute (3.26) with

$$\mathcal{V}_2 \quad \geq \quad \frac{1}{T-2}\sum_{t=2}^{T-1}[(y_{t+1}^x - y_t^x) - (y_t^x - y_{t-1}^x)]^2 \tag{3.27}$$

Note that if $\mathcal{V}_2$ is the same across units, the acceleration in the trend is common. Therefore (3.27) imposes some form of balance growth across countries. The main difference between (3.26) and (3.27) is that the former allows countries with more volatile cyclical component to have also more volatile trend, while this is not possible in the latter.

Endogenously selecting the frequencies belonging to the cycle can be useful in certain contexts but care should be exercised since uncritical application of this idea may lead to absurd conclusions if the mechanism generating the data differs across units.

Exercise 3.26 *Consider the following processes: i)*$(1-0.99\ell)y_t = e_t$; *ii)*$(1-1.34\ell+0.7\ell^2)y_t = e_t$; $y_t = (1-0.99\ell)e_t$. *Show the implied value of* λ *in the three cases when* $\mathcal{V}_1 = 0.5$ *and* $\bar{\lambda} = 1600$ *and the resulting business cycle frequencies.*

3.2.2 Exponential smoothing (ES) filter

The exponential smoothing filter, used e.g. in Lucas (1980), is obtained from the program:

$$\min_{y_t^x}\{\sum_{t=0}^{T}(y_t - y_t^x)^2 + \lambda\sum_{t=0}^{T}(y_t^x - y_{t-1}^x)^2\} \tag{3.28}$$

The ES filter therefore differs from the HP filter in the penalty function: here we penalize changes in the trend, while in the HP we penalize the acceleration of the trend.

The first order conditions of the problem are $0 = -2(y_t - y_t^x) + 2\lambda(y_t^x - y_{t-1}^x) - 2\lambda(y_{t+1}^x - y_t^x)$. Therefore, as in the HP filter, the trend component is $y_t^x = (\mathbb{F}^{ES})^{-1}y_t$ and the cyclical component is $y_t^c = y_t - y_t^x = (1-(\mathbb{F}^{ES})^{-1})y_t$.

Exercise 3.27 *Display the form of* $\mathbb{F}^{ES}$ *and compare it with* $\mathbb{F}^{HP}$.

Note that if t runs from $-\infty$ to ∞, the solution to the minimization problem can be written as $y_t^x = [\lambda(1-\ell)(1-\ell^{-1})+1]^{-1}y_t$.

Example 3.16 *The ES filter automatically removes a linear trend from* y_t. *To show this let* $\mathbb{F}^{ES}y_t^x = y_t$ *and* $\mathbb{F}^{ES}\tilde{y}_t^x = \tilde{y}_t$, *where* $\tilde{y}_t = y_t + a_0 + a_1 t$. *Combining the two expressions we have* $\mathbb{F}^{ES}(y_t^x - \tilde{y}_t^x) = y_t - \tilde{y}_t = -a_0 - a_1 t$ *or* $\mathbb{F}^{ES}(y_t^c - \tilde{y}_t^c) + (\mathbb{F}^{ES} - 1)(-a_0 - a_1 t) = 0$. *Hence for* $y_t^c = \tilde{y}_t^c$ *we need* $(\mathbb{F}^{ES} - 1)(-a_0 - a_1 t) = 0$. *The result follows since* $(\mathbb{F}^{ES} - 1)$ *is symmetric.*

Exercise 3.28 *Using the same logic of example 3.16 examine whether the ES filter is able to remove a quadratic trend from the data.*

Given the form of the trend remover ES filter, one can show that $y_t^c = (1 - \mathcal{B}^x(\ell))y_t = \frac{(1-\ell)(1-\ell^{-1})}{\frac{1}{\lambda}+(1-\ell)(1-\ell^{-1})}y_t$. Hence, application of the ES filter induces stationarity in y_t^c for y_t integrated up to order 2. Conversely, if y_t is integrated of order less than 2, y_t^c will display unit root moving average and therefore strong (and possibly artificial) persistence.

The next exercise shows the effect of applying the ES filter to various types of data.

Exercise 3.29 *i) Show that when* y_t *is stationary, the gain function of the cyclical ES filter is* $\frac{2(1-\cos(\omega))}{\frac{1}{\lambda}+2(1-\cos(\omega))}$. *Show that* y_t^c *has zero power at* $\omega = 0$, *has the same power as* y_t *at* $\omega \to \pi$ *and that the larger is* λ, *the smoother is* y_t^x *(the more variable is* y_t^c*).*
ii) Show that when y_t *is I(2), the gain function of the cyclical filter is* $\frac{1}{\frac{1}{\lambda}+2(1-\cos(\omega))}$. *Describe what is the effect of this filter at* $\omega = 0, \pi$ *and at business cycle frequencies.*

The broad similarities of ES and HP filter can be appreciated in figure 3.2 where we plot the gain function of the two filters when $\lambda = 1600$. It is clear that the ES filter picks up trends with longer periodicity, but generally speaking, the two filters are very similar.

Exercise 3.30 *Let* $y_t = \rho_y y_{t-1} + e_t$, $e_t \sim$ *iid* $(0,1)$ *and* $\rho_y = 0.5, 0.9, 1.0$. *Simulate 2000 data points with* $y_0 = 10$ *and pass the last 1500 through the HP and the ES filters. Plot the cyclical components, compute their variability and their auto and cross correlation.*

Both the HP and the ES trend extractors are special cases of a general class of low pass filters that engineers call Butterworth (BW) filters. Such filters have a squared gain function of the form $|Ga(\omega)|^2 = 1/(1 + (\frac{\sin(\omega/2)}{\sin(\bar{\omega}/2)})^{2\kappa})$ where κ is a parameter and $\bar{\omega}$ is the frequency where the frequency response of the filter is equal to 0.5. For BW filters, the trend estimate is $y_t^x = \frac{1}{1-\lambda(1-\ell)^{\kappa}(1-\ell^{-1})^{\kappa}}y_t$ where $\lambda = \frac{1}{2^{2\kappa}sin^{2\kappa}(\bar{\omega}/2)}$.

Example 3.17 *It is easy to show that if* $\kappa = 2$ *and* $\bar{\omega}$ *solves* $\lambda = (16\sin^4(\bar{\omega}/2))^{-1}$, $|Ga(\omega)|^2$ *is the square gain function of the HP trend extractor filter while if* $\kappa = 1$ *and* $\bar{\omega}$ *solves* $\lambda = (4\sin^2(\bar{\omega}/2))^{-1}$, *it produces the square gain function of the ES trend extractor filter.*

Relative to HP and ES filters, general BW filters have a free parameter κ, which can be used to tailor the gain function to particular needs. In fact, a higher κ moves $|Ga(\omega)|^2$ to the right (i.e. $\bar{\omega}$ increases). Hence, for a fixed λ, it controls which cycles are included in y_t^c. Designing a low pass BW filter is easy. We need two parameters a_1, a_2 and two frequencies ω_1, ω_2 such that $1 - a_1 < Ga^{BW}(\omega) \leq 1$ for $\omega \in (0, \omega_1)$ and $0 < Ga^{BW}(\omega) \leq a_2$ for $\omega \in (\omega_2, \pi)$. Given, $a_1, a_2, \omega_1, \omega_2$, one finds $\bar{\omega}$ and κ solving $1 + (\frac{\sin(\omega_1/2)}{\sin(\bar{\omega}/2)})^{2\kappa} = (1 - a_1)^{-1}$ and $1 + (\frac{\sin(\omega_2/2)}{\sin(\bar{\omega}/2)})^{2\kappa} = (a_2)^{-1}$, rounding off κ to the closest integer.

3.2.3 Moving average (MA) filters

MA filters have a long history as smoothing devices and their use goes back, at least, to the work of Burns and Mitchell (1946). MA filters are defined by a polynomial in the lag operator $\mathcal{B}(\ell)$ which is either one or two-sided (that is, it operates on J lags or on J leads and J lags of y_t). A MA filter is symmetric if $\mathcal{B}_j = \mathcal{B}_{-j}, \forall j$.

The frequency response function of a symmetric MA filter is $\mathcal{B}(\omega) = \mathcal{B}_0 + 2\sum_j \mathcal{B}_j \cos(\omega j)$, where we have used the trigonometric identity $2\cos(\omega) = \exp(i\omega) + \exp(-i\omega)$. Symmetric filters are typically preferred since they have zero phase shift. This is a desirable property since for $Ph(\omega) = 0,\ \forall\omega$, the timing of the cycles in y_t and $\mathcal{B}(\ell)y_t$ is the same.

Example 3.18 *A simple symmetric two-sided (truncated) moving average filter is* $\mathcal{B}_j = \frac{1}{2J+1},\ 0 \leq j \leq |J|$ *and* $\mathcal{B}_j = 0, j > |J|$. *If we set* $y_t^c = (1 - \mathcal{B}(\ell))y_t \equiv \mathcal{B}^c(\ell)y_t$ *the cyclical weights are* $\mathcal{B}_0^c = 1 - \frac{1}{2J+1}$ *and* $\mathcal{B}_j^c = \mathcal{B}_{-j}^c = -\frac{1}{2J+1}$, $j = 1, 2 \ldots, J$. *It is easy to recognize that* $\mathcal{B}_j$ *are the weights used in the Box-car kernel in chapter 1. The Bartlett and the quadratic spectral kernels are also two-sided symmetric filters.*

Note that $\mathcal{B}(\omega = 0) = \sum_{j=-\infty}^{\infty} \mathcal{B}_j$. Therefore, the condition $\lim_{J\to\infty} \sum_{j=-J}^{J} \mathcal{B}_j = 1$ is necessary and sufficient for a MA filter to have unitary gain at the zero frequency. If this is the case, $\mathcal{B}^c(\omega = 0) = 1 - \mathcal{B}(\omega = 0) = 0$, and y_t^c has zero power at the zero frequency.

Example 3.19 *The effect of asymmetric filters can be appreciated in figure 3.4 where we present the gain of the filter of example 3.18, of the HP filter, and of an asymmetric filter with right side equal to the one of example 3.18 and left side with weights* $\frac{1}{2j+1},\ j < J = 12$. *In general, MA filters have unit gain only for* $\omega \approx \pi$ *and leave a lot of high frequency variability in the trend. Relative to a symmetric MA filter, an asymmetric one has gain different from zero at* $\omega = 0$ *and leaves much more cyclical variability in the trend.*

Exercise 3.31 *(Baxter and King) i) Show that a symmetric MA filter with* $\lim_{J\to\infty} \sum_{-J}^{J} \mathcal{B}_j = 1$ *is sufficient to extract the quadratic trend from* $y_t = e_t + a_0 + a_1 t + a_2 t^2$, *where* e_t *is arbitrarily serially correlated but stationary process.*
ii) Show that if $\lim_{J\to\infty} \sum_{-J}^{J} \mathcal{B}_j = 1$, *the cyclical filter can be decomposed as* $\mathcal{B}^c(\ell) = (1 - \ell)(1 - \ell^{-1})\mathcal{B}^{c\dagger}(\ell)$, *where* $\mathcal{B}^{c\dagger}$ *is a symmetric MA filter with* $J - 1$ *leads and lags. That is,* y_t^c *will be stationary when* y_t *integrated of order up to two.*

Exercise 3.32 *Is the ES filter a symmetric MA filter? Does it satisfy* $\lim_{J\to\infty}\sum_{j=J}^{J}\mathcal{B}_j = 1$*?*

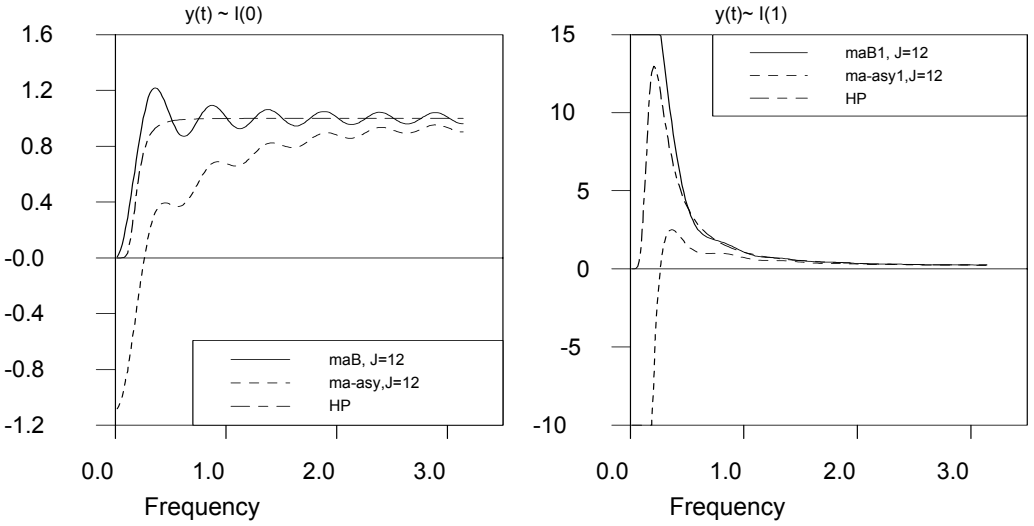


Figure 3.4: Gain: symmetric and asymmetric MA and HP filters

Example 3.20 *One type of MA filter extensively used in the seasonal adjustment literature is the so-called Henderson filter. The filter is symmetric, operates on* J *leads and* J *lags of* y_t *and* $\mathcal{B}_j$ *are found solving* $\min_{j=-J}^{J}((1-\ell)^3\mathcal{B}_j)^2$ *subject* $\sum_j \mathcal{B}_j = 1, \sum j\mathcal{B}_j = 0, \sum_j j^2\mathcal{B}_j = 0$*. Intuitively, the objective function of the problem measures the degree of smoothness of the curve described by the weights. The constraints imply that polynomials of degree up to the second are required to be part of the weights. When* $J = 6$*,* $\mathcal{B}_0 = .2401$ *and* $\mathcal{B}_j = (.2143, .1474, .0655, 0, -0.279, -0.19)$*. These bell-shaped weights define a filter whose gain function resembles the one of the HP trend extractor filter and is smoother than the one constructed using a tent-like filter.*

3.2.4 Band Pass (BP) filters

Band Pass filters have become popular in applied macro following the work of Canova (1998), Baxter and King (1999) and Christiano and Fitzgerald (2003). One reason for preferring BP filters is that the majority of the other filters have high pass characteristics and therefore leave or exaggerate the amount of variability present at high frequencies. As we have seen in chapter 1, band pass filters are combinations of MA filters designed to eliminate both

high and low frequencies movements in the data. Furthermore, BP filters are appealing because they make the notion of business cycle operational by selecting fluctuations in a prespecified range (say, 6 to 24 quarters).

The output of high, low and band pass filters can be represented in time domain with infinite two-sided symmetric moving averages of y_t. In chapter 1 we have seen that the coefficients of a low pass filter are $\mathcal{B}_0^{lp} = \frac{\omega_1}{\pi}$; $\mathcal{B}_j^{lp} = \frac{sin(j\omega_1)}{j\pi}$, $j = \pm 1, \pm 2, \ldots$ where ω_1 is the upper frequency of the band; the ones of a high pass filter are $\mathcal{B}_0^{hp} = 1 - \mathcal{B}_0^{lp}$; $\mathcal{B}_j^{hp} = -\mathcal{B}_j^{lp}$ and those of a band pass filter are $\mathcal{B}_j^{bp} = \mathcal{B}_j^{lp}(\omega_2) - \mathcal{B}_j^{lp}(\omega_1)$ for $\omega_1 < \omega_2$. Unfortunately, with a finite amount of data, these filters are not implementable. Therefore, one needs to approximate them with finite MA weights. One set of approximating weights can be found truncating optimal ones.

Exercise 3.33 *(Koopman) Show that if one chooses the finite symmetric approximation $\mathcal{B}_J^A$ that minimizes $\int_\pi^\pi |\mathcal{B}^A(\omega) - \mathcal{B}(\omega)|^2 d\omega$, where J stands for the number of leads/lags used and $\mathcal{B}(\omega)$ is the ideal filter, the solution is $\mathcal{B}_j^A = \mathcal{B}_j$ for $|j| \leq J$ and $\mathcal{B}_j^A = 0$ otherwise.*

Intuitively, such a truncation is optimal since the weights for $|j| > J$ are small.

To insure that the approximating BP filter has unit root removal properties we impose $\mathcal{B}^A(\omega = 0) = 0$. The next exercise shows how to modify $\mathcal{B}_J^A$ to account for this restriction.

Exercise 3.34 *(Baxter and King) Show that for a low pass filter, imposing $\mathcal{B}^A(\omega = 0) = 1$ implies that the constrained approximate weights are $\mathcal{B}_j^A + \frac{1 - \sum_{j=-J}^{J} \mathcal{B}_j^A}{2J+1}$. Show the constrained approximate weights for the approximate BP filter.*

Clearly, the quality of the approximation depends on the truncation point J (see chapter 5 for a similar problem). One way to quantify the biases introduced by the truncation is the following.

Exercise 3.35 *i) Plot the gain function of the optimal and the approximate band pass filter obtained when J=4,8,12,24 Examine both the leakage and the compression that the approximate filter has relative to the optimal one at business cycle frequencies.*
ii) Simulate $y_t = 0.9y_{t-1} + e_t$, where $e_t \sim$ iid $\mathbb{N}(0,1)$. Apply the approximate band pass filter weights for $J = 4, 8, 12, 24$. Calculate sample statistics of the filtered data for each J.
iii) Simulate the model $(1 - \ell)y_t = e_t - 0.7e_{t-1}$. Repeat the steps in point ii).

Roughly speaking, J must be sufficiently large for the approximation to be reasonable. However, the larger is J, the shorter is the time series for y_t^c available (we are loosing J observations at the beginning and at the end of the sample), and therefore the less useful the approximation is to measure the current state of the cycle. Simulation studies have shown that if one is to extract cycles with periodicity between 6-24 quarters, a constrained band pass filter with $J \approx 12$ has little leakage and minor compression relative to other filters and produces cyclical components which are similar (but less volatile) to those extracted with HP filter for observations to the middle of the sample.

The modified approximate band-pass filter has the same problems as other high pass filters when applied to I(1) series. In fact, for symmetric MA filters with zero gain at $\omega = 0$ we can write $\mathcal{B}^A(\ell) = -(1-\ell)(1-\ell^{-1})\mathcal{B}^{A*}(\ell)$, where $\mathcal{B}^{A*}_{|j|} = \sum_{i=|j|+1}^{J}(i-|j|)\mathcal{B}^A_i$.

The next exercise shows that if the data are stationary up to a quadratic trend, then no distortion in y^c_t results. Distortions however obtain if y_t is integrated.

Exercise 3.36 *(Murray) Show that if $y_t = a + b_1 t + b_2 t^2 + e_t$, then $y^c_t = \mathcal{B}^A(\ell)e_t$. Show that if $(1-\ell)y_t = e_t$, then $y^c_t = -(1-\ell^{-1})\bar{B}^{A*}(\ell)e_t$. Plot the gain functions of $\mathcal{B}^A(\ell)$ and $-(1-\ell^{-1})\mathcal{B}^{A*}(\ell)$ and describe their differences.*

Example 3.21 *If y_t is integrated, the BP filter may also generate spurious periodicity in filtered data. To show this we have generated 1000 samples of 500 data points from $\Delta y_t = e_t,\ e_t \sim\ iid\ (0,\sigma^2)$; and constructed y^c_t using the $-(1-\ell^{-1})\mathcal{B}^{A*}(\ell)$ filter. We then computed the mean value of the ACF of y^c_t for $J = 4, 8, 12$. The ACF of Δy_t is zero for all $\tau \geq 1$ while $ACF_{y^c}(\tau)$ is different from zero, at least for $\tau < 10$. Hence an integrated process produces autocorrelated y^c_t if passed with the above filter. Interestingly, the persistence of y^c_t increases with J. For example, the mean of $ACF_{y^c}(\tau)$ fails to converge to zero for $J = 12$ for at least $\tau \leq 15$.*

Exercise 3.37 *(Murray) Let $y_t = y^x_t + y^c_t$ and let*

$$y^x_t = 0.82 + y^x_{t-1} + e^x_t \quad e^x_t \sim\ iid\ \mathbb{N}(0,(1.24)^2) \tag{3.29}$$

$$(1 - 1.34\ell + 0.71\ell^2)y^c_t = e^c_t \quad e^c_t \sim\ iid\ \mathbb{N}(0,(0.75)^2) \tag{3.30}$$

i) Calculate the autocovariance function of the cyclical component.
ii) Simulate 2000 data points for y_t, filter them with an approximate BP filter using J=8, 16, 24, 40. Calculate the autocovariance function of the estimated y^c_t.
iii) Simulate 2000 data points for y_t setting $var(e^x_t) = 0$. Pass the simulated time series through an approximate BP filter using J=8,16,24,40. Calculate the autocovariance function of the estimated cyclical component. When is the autocovariance function calculated in ii) and iii) closer to the one you have calculated in i)?
iv) Repeat i), ii) and iii) 1000 times and store the values of the first five elements of the ACF. Calculate the number of times that the ACF in i) lies within the 68% band you have computed for each step.

The approximate BP filter of exercise 3.33 is constructed equally penalizing deviations from the ideal filter at all frequencies. This may not be the best approximating distance. Intuitively, we would like the approximate filter to reproduce as closely as possible the ideal filter at those frequencies where the spectrum of y_t is large while we are less concerned about deviations when the spectrum of y_t is small. Christiano and Fitzgerald (CF) (2003) construct an approximation to the ideal filter which has these features using projection techniques. The filter they obtain is non-stationary, asymmetric, and depends on the time series properties of y_t. The non-stationarity comes from the fact that there is a different

projection problem for each t. Asymmetry is produced since all observations are used at each t to construct the filtered series. The dependence on the properties of y_t comes from the fact that the power of $\mathcal{S}_y(\omega)$ at different ω depends on the features of y_t. Contrary to the approximating filter of exercise 3.33, the approximating filter of Christiano and Fitzgerald does not truncate the optimal weights, except for some special DGP. Note that the CF filter could be made stationary and symmetric if these features are deemed necessary.

The CF filter can be obtained as follows. Suppose we want to choose $\mathcal{B}_j^{A,t-1,T-t}$, $j = T-t,\ldots,t-1$, to minimize $\int_{-\pi}^{\pi} \frac{|\mathcal{B}^{A,t-1,T-t}(\omega)-\mathcal{B}(\omega))|}{1-e^{-i\omega}}\mathcal{S}_{\Delta y}(\omega)d\omega$. CF show that the solution to this problem can be represented as a $(T+1)$ system of linear equations of the form $\mathbb{F}_0^{CF} = \mathbb{F}_1^{CF}\mathcal{B}^{A,t-1,T-t}$ where $\mathbb{F}_0^{CF}, \mathbb{F}_1^{CF}$ depend on the properties of $\mathcal{S}_{\Delta y}$ and $\mathcal{B}(\omega)$.

Example 3.22 *There are few cases for which the solution to the problem is of interest. The first is when y_t is a random walk. Here, the approximate band pass filtered version of y_t is $y_t^c = \mathcal{B}_{T-t}^A y_T + \mathcal{B}_{T-t-1}^A y_{T-1} + \ldots + \mathcal{B}_1^A y_{t+1} + \mathcal{B}_0^A y_t + \mathcal{B}_1^A y_{t-1} + \ldots + \mathcal{B}_{t-2}^A y_2 + \mathcal{B}_{t-1}^A y_1$, for $t = 2,3,\ldots T-1$, where $\mathcal{B}_0^A = \frac{\frac{2\pi}{\omega_1}-\frac{2\pi}{\omega_2}}{\pi}$; $\mathcal{B}_j^A = \frac{\sin(2j\pi/\omega_1)-\sin(2j\pi/\omega_2)}{j\pi}$, $j \neq t-1, T-t$; $\mathcal{B}_{T-t}^A = -0.5\mathcal{B}_0^A - \sum_{j=1}^{T-t-1}\mathcal{B}_j^A$ while $\mathcal{B}_{t-1}^A$ solves $0 = \mathcal{B}_0^A + \mathcal{B}_1^A + \ldots + \mathcal{B}_{T-1-t}^A + \mathcal{B}_{T-t}^A + \mathcal{B}_1^A + \ldots + \mathcal{B}_{t-2}^A + \mathcal{B}_{t-1}^A$, where $\omega_2(\omega_1)$ is the upper (lower) frequency of the band. For $t = 1$ the expression is $y_1^c = 0.5\mathcal{B}_0^A y_1 + \mathcal{B}_1^A y_2 + \ldots + \mathcal{B}_{T-2}^A y_{T-1} + \mathcal{B}_{T-1}^A y_T$ and for $t = T$ is $y_T^c = 0.5\mathcal{B}_0^A y_T + \mathcal{B}_1^A y_{T-1} + \ldots + \mathcal{B}_{T-2}^A y_2 + \mathcal{B}_{T-1}^A y_1$. From the above, it is clear that the filter uses all the observations for each t, that the weights change with t and become asymmetric if t is away from the middle of the sample.*

A second interesting case obtains if y_t is iid in which case the weights $\mathcal{S}_y(\omega) = \frac{\sigma_y^2}{2\pi}$ are independent of ω. Then $\mathcal{B}_j^{A,t-1,T-t} = \mathcal{B}_j$ for $j = T-t,\ldots,t-1$ and zero otherwise, which produces the solution of exercise 3.33 if the filter is required to be symmetric and $\mathcal{B}_j$ is truncated for $j > J$.

We present the gain function of the approximate asymmetric and approximate symmetric truncated CF filters, both obtained when y_t is a random walk, of the ideal filter and of the approximate BK filter produced in exercise 3.33 in figure 3.5. The two truncated symmetric filters are similar but the CF filter gives more weights to the lower frequencies of the band (since a random walk has more power in those frequencies) and has smaller side loops. The asymmetric filter, on the other hand, is very close to the ideal one.

Exercise 3.38 *Describe how to make the approximate CF filter $\mathcal{B}^A(\ell)$ symmetric.*

The approximate CF filter is general and solves beginning and end-of-the-sample problems. The costs of this general setup is that one has to decide a-priori how $\mathcal{S}_{\Delta y}(\omega)$ looks like (in particular, if y_t is stationary or integrated and what are its serial correlation properties) and that the filter imposes phase shifts in the autocovariance function of y_t. By way of examples, Christiano and Fitzgerald suggest that phase shifts are small in practice and can be safety neglected in the analysis. Also, they indicate that, in practice, the approximation obtained arbitrarily assuming that y_t is a random walk works well for a variety of macroeconomic time series.

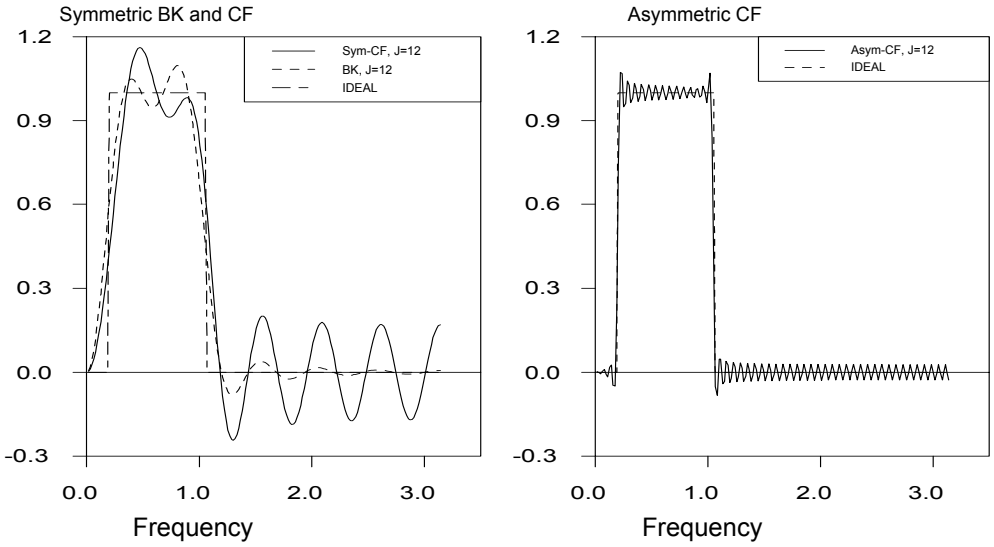


Figure 3.5: Gain, ideal and approximate BP filters

As other MA filters, the approximate CF filter also faces problems when y_t is integrated. In fact, it increases the variability at business cycle frequencies if a term $(1-\ell)$ is used to make y_t stationary.

What kind of statistics should one use to compare the quality of the approximation to the ideal BP filter? The following exercise suggests two possible alternatives.

Exercise 3.39 *Suppose the optimal band pass filtered series is y_t^c and the approximate band pass filtered series is y_t^{Ac}. Let $y_t^c = y_t^{Ac} + e_t$, where by optimality of the projection problem $E(e_t|\mathcal{F}_t) = 0$, and $\mathcal{F}_t$ is the information set at time t. Show that $var(y_t^c - y_t^{Ac}|\mathcal{F}_t) = var(y_t^c)(1 - corr(y_t^{Ac}, y_t^c))$. Conclude that $corr(y_t^{Ac}, y_t^c)$ and $\frac{var(y_t^{Ac})}{var(y_t^c)}$ can be used to evaluate the closeness of the approximation.*

Exercise 3.40 *Consider the DGP used in exercise 3.37. Simulate data for the two components, compute y_t 1000 times and calculate $ACF(\tau)$ of y_t^c for $\tau = 1, \ldots 6$. For each draw, estimate y_t^c using the fixed weight approximate BP filter, the non-stationary, asymmetric BP filter and the simulated y_t, and compute $ACF(\tau)$ of y_t^c for $\tau = 1, \ldots 6$. Using the true and the simulated distribution of $ACF(\tau)$ for each BP filter, examine which method better approximates the cyclical component of the data.*

While it has become common to use the time domain representation of the filter and therefore worry about the effects of truncation, one can directly implement BP filters in

frequency domain (see Canova (1998)). The advantage of this approach is that no approximation is needed and no loss of data is involved. However, two major drawbacks need to be mentioned. First, the definition of the cyclical component depends on the sample size. This is because Fourier frequencies are function of T. Hence, when new information arrives, the measurement of y_t^c for all t needs to be changed. The time domain version of the truncated BP filter does not have this problem since the filter weights are independent of t. Second, since the spectrum of y_t is undefined at $\omega = 0$ when the series is non-stationary, a stationary transformation is required before the spectrum is computed. Hence one should decide whether a deterministic or a stochastic trend should be preliminary removed.

Recently Corbae and Ouliaris (2001) suggested a way to implement band pass filters in frequency domain which does not suffer from the latter problem. Their implementation is useful since it also solves the problem of the spurious periodicity induced by band pass filters when y_t is integrated. Suppose $\Delta y_t = D(\ell)e_t$ where $e_t \sim (0, \sigma^2)$ with finite fourth moments and $\sum_j D_j^2 < \infty$. For $\omega \neq 0$ Corbae, Ouliaris and Phillips (COP)(2002) showed that the spectrum of y_t is

$$\mathcal{S}_y(\omega) = \frac{1}{1-e^{i\omega}} D(\omega)\mathcal{S}_e(\omega) - \frac{1}{\sqrt{T}} \frac{e^{i\omega}}{1-e^{i\omega}} (y_T - y_0) \tag{3.31}$$

where the last term is the bias induced by the unit root at $\omega = 0$. From (3.31) one can see that $\mathcal{S}_y(\omega_1)$ is not independent of $\mathcal{S}_y(\omega_2)$ for $\omega_1 \neq \omega_2$ and both Fourier frequencies. Since $(y_T - y_0)$ is independent of ω and does not disappear even if $T \to \infty$, COP also show that the leakage from frequency zero creates biases in y_t^c that do not disappear if y_t is first detrended in time domain.

(3.31) suggests a simple way to eliminate this bias. The last term looks like a deterministic trend (in frequency domain) with random coefficient $(y_T - y_0)$. Hence, to construct an ideal BP filter in frequency domain when y_t is integrated, one could use the following:

Algorithm 3.2

1) *Compute* $\mathcal{S}_y(\omega)$ *for* $\omega \neq 0$.

2) *Run a (cross-frequency) regression of* $\mathcal{S}_y(\omega)$ *on* $\frac{1}{\sqrt{T}} \frac{e^{i\omega}}{1-e^{i\omega}}$ *for* $\omega \in (0, \pi]$ *and let* $\widehat{(y_T - y_0)}$ *be the resulting estimator of* $(y_T - y_0)$.

3) *Construct* $\mathcal{S}_y^*(\omega) = \mathcal{S}_y(\omega) - \widehat{(y_T - y_0)} \frac{1}{\sqrt{T}} \frac{e^{i\omega}}{1-e^{i\omega}}$. *Apply the ideal band pass filter to* $\mathcal{S}_y^*(\omega)$.

Two features of algorithm 3.2 should be mentioned. First $\widehat{y_T - y_0}$ is a $\sqrt{T}$- consistent estimator of $(y_T - y_0)$ and when y_t is stationary $\widehat{y_T - y_0} = 0$. Second, no data is lost because of the filter and no parameters are chosen by the investigator (the approximate BP filter requires, at least, a truncation point J).

Exercise 3.41 *Suppose you are willing to loose two observations,* y_T *and* y_0. *Show how to modify algorithm 3.2 to obtain an ideal BP filter. Intuitively describe why this modified filter can have better finite sample properties then the one produced by algorithm 3.2.*

3.3 Economic Decompositions

The decompositions of this group are diverse, but have one feature in common: they all use an economic model to guide the extraction of the cyclical component. They should be more appropriately called permanent-transitory decompositions since they define the trend as the component of the series driven by permanent shocks. All decompositions use structural VARs (which we discuss in chapter 4) even though the "level" of identification is minimal. In fact, instead of trying to obtain behavioral disturbances, they simply look for shocks with permanent or transitory features.

3.3.1 Blanchard and Quah (BQ) Decomposition

The most prominent decomposition of this class was suggested by Blanchard and Quah (1989) who use a version of Fischer (1979) partial equilibrium model with overlapping labor contracts consisting of four equations:

$$\begin{aligned} GDP_t &= M_t - P_t + a\zeta_t && (3.32)\\ GDP_t &= N_t + \zeta_t && (3.33)\\ P_t &= W_t - \zeta_t && (3.34)\\ W_t &= W|\{E_{t-1}N_t = N^{fe}\} && (3.35) \end{aligned}$$

where $M_t = M_{t-1} + \epsilon_{3t}$, $\ln \epsilon_{3t} \sim \ iid\ (0, \sigma_M^2)$, $\zeta_t = \zeta_{t-1} + \epsilon_{1t}$, $\ln \epsilon_{1t} \sim \ iid\ (0, \sigma_\zeta^2)$ and where GDP_t is output, N_t is employment, N^{fe} is full employment, M_t is money, P_t are prices, ζ_t is a productivity disturbance - all these variables are measured in logs - and W_t is the real wage. The first equation is an aggregate demand equation, the second a short run production function, the third and the fourth describe price and wage setting behavior. Here money supply and productivity are exogenous and integrated processes. Also, contrary to the models of chapter 2, these equations are postulated and not derived from micro-principles.

Letting $UN_t = N_t - N^{fe}$, the solution to the model implies a bivariate representation for (GDP_t, UN_t) of the form

$$\begin{aligned} GDP_t &= GDP_{t-1} + \epsilon_{3t} - \epsilon_{3t-1} + a(\epsilon_{1t} - \epsilon_{1t-1}) + \epsilon_{1t} && (3.36)\\ UN_t &= -\epsilon_{3t} - a\epsilon_{1t} && (3.37) \end{aligned}$$

The model therefore places restrictions on the data. In particular (3.36)-(3.37) imply that fluctuations in UN_t are stationary while GDP_t is integrated. Furthermore, its permanent component is $GDP_t^x \equiv GDP_{t-1} + a(\epsilon_{1t} - \epsilon_{1t-1}) + \epsilon_{1t}$ and its transitory component is $GDP_t^c \equiv \epsilon_{3t} - \epsilon_{3t-1}$. In other words, while demand shocks drive GDP cycle, both supply and demand shocks drive the cycle in UN_t. To extract the transitory component of GDP we need the following steps:

Algorithm 3.3

1) *Check that GDP_t is integrated and UN_t is stationary (possibly after some transformation). This step is necessary for the decomposition to be meaningful.*

2) *Identify two shocks, one which has a permanent effect on GDP_t and one which has a transitory effect on both GDP_t and UN_t.*

3) *Compute $GDP_t^c = GDP_t - GDP_t^x$ and $UN_t^c \equiv UN_t$.*

In step 2) one could generically specify a bivariate VAR for the data (if the model is believed to provide only qualitative restrictions) or condition on the exact structure provided by (3.36)-(3.37) to derive the shocks.

It is important to stress that decompositions like (3.36)-(3.37) are conditional on the economic model. Hence, it is possible to produce different cyclical components using the same model but introducing different features or frictions.

Exercise 3.42 *(Lippi and Reichlin) Suppose the productivity shock has the structure $\zeta_t = \zeta_{t-1} + \mathbb{Q}(\ell)\epsilon_{1t}$ where $\sum_j \mathbb{Q}_j = 1$ and $\ln \epsilon_{1t} \sim iid(0, \sigma^2)$. Show that a solution for ΔGDP_t and UN_t can be written as $\begin{bmatrix} \Delta GDP_t \\ UN_t \end{bmatrix} = \begin{bmatrix} 1-\ell & (1-\ell)a + \mathbb{Q}(\ell) \\ -1 & -a \end{bmatrix} \begin{bmatrix} \epsilon_{1t} \\ \epsilon_{3t} \end{bmatrix}$. Argue that trend and cycle may not be identifiable from the data for $y = [\Delta GDP_t, UN_t]$. (Hint: the data has a representation $y_t = \bar{y} + D(\ell)e_t$, where the roots of $D(\ell)$ are on or outside the complex unit circle, see also section 6 of chapter 4).*

Exercise 3.43 *(Gali) Suppose a representative household maximizes $E_0 \sum_t \beta^t [\ln C_t + \vartheta_M \ln(\frac{M_t}{P_t}) - \frac{\vartheta_n}{1-\varphi_n} N_t^{1-\varphi_n} - \frac{\vartheta_{ef}}{1-\varphi_{ef}} ef_t^{1-\varphi_{ef}}]$, where $C_t = (\int_0^1 C_{it}^{\frac{1}{1+\varsigma_p}} di)^{1+\varsigma_p}$, ς_p is the elasticity of substitution, $p_t = (\int_0^1 p_{it}^{-\frac{1}{\varsigma_p}} di)^{-\varsigma_p}$ is the aggregate price index, $\frac{M_t}{P_t}$ are real balances, N_t is hours worked and ef_t is effort. The budget constraint is $\int_0^1 p_{it} C_{it} di + M_t = w_{Nt} N_t + w_{et} ef_t + M_{t-1} + T_t + Prf_t$, where T_t are monetary transfers and Prf_t profits. A continuum of firms produces a differentiated good: $inty_{it} = \zeta_t (N_{it}^{\eta_2} ef_{it}^{1-\eta_2})^{\eta_1}$, where $N_{it}^{\eta_2} ef^{1-\eta_2}$ is the quantity of effective input, ζ_t an aggregate technology shock, $\Delta\zeta_t = \epsilon_{1t}$ where $\ln \epsilon_{1t} \sim$ iid $(0, \sigma_\zeta^2)$. Firms set prices one period in advance, taking p_t as given but not knowing the current realization of the shocks. Once shocks are realized, firms optimally choose employment and effort. So long as marginal costs are below the predetermined price, firms will then meet demand and choose an output level, equal to $(\frac{p_{it}}{p_t})^{-\frac{1+\varsigma_p}{\varsigma_p}} C_t$. Optimal price setting implies $E_{t-1}[\frac{1}{C_t}((\eta_1\eta_2) p_{it} inty_{it} - (\varsigma_p + 1) w_{Nt} N_{it})] = 0$. Assume $\Delta M_t = \epsilon_{3t} + a_M \epsilon_{1t}$, where $\ln \epsilon_{3t} \sim$ iid $(0, \sigma_M^2)$ and a_M is a parameter.*
i) Letting lower case letters denote natural logs, show that, in equilibrium, output growth (Δgdp), log employment (n_t), and labor productivity growth (Δnp) satisfy

$$\Delta gdp_t = \Delta\epsilon_{3t} + a_M \epsilon_{1t} + (1 - a_M)\epsilon_{1t-1} \tag{3.38}$$

$$n_t = \frac{1}{\eta_s}\epsilon_{3t} - \frac{1-a_M}{\eta_s}\epsilon_{1t} \tag{3.39}$$

$$\Delta np_t = (1-\frac{1}{\eta_s})\Delta\epsilon_{3t} + (\frac{1-a_M}{\eta_s} + a_M)\epsilon_{1t} + (1-a_M)(1-\frac{1}{\eta_s})\epsilon_{1t-1} \tag{3.40}$$

where $np_t = gdp_t - n_t$ *and* $\eta_s = \eta_1(\eta_2 + (1-\eta_2)\frac{1+\varphi_n}{1+\varphi_{ef}})$.
ii) Describe a trend-cycle decomposition using $(\Delta gdp, n_t)$. *How does this decomposition differ from the one computed using* $(\Delta np_t, n_t)$*?*

BQ and multivariate BN decompositions share important similarities. In both cases, in fact, the trend is a random walk. However, while here trend and cycle are driven by orthogonal shocks, in the BN decomposition they are driven by the same combination of shocks. Hence, the disturbances in a BQ decomposition have some vague economic interpretation, while this is not the case for those of a multivariate BN decomposition.

Example 3.23 *We have derived BQ and BN decompositions of a bivariate system with* ΔGDP_t *and* UN_t *(proxied by the unemployment rate) using US data for the period 1950:1-2003:3. Both series are demeaned and a linear trend is eliminated from the unemployment rate. The cyclical component of output is computed using "structural" shocks (BQ decomposition) or "reduced form" shocks (BN decomposition). The estimated cyclical components are quite different. For example, while both of them have similar AR(1) coefficient (0.93 for BN and 0.90 for BQ), their contemporaneous correlation is only 0.21. This occurs because the BQ cyclical component is much more volatile (the standard error is 2.79 as opposed to 0.02) and that the swings induced by temporary shocks have longer duration - about 10 quarters while the mean length of BN cycles is about 5 quarters.*

3.3.2 King, Plosser Stock and Watson (KPSW) Decomposition

King, Plosser, Stock and Watson (1991) start from a RBC model where the log of total factor productivity (TFP) is driven by a unit root. This assumption implies that all endogenous variables but employment will be trending and that the trend will be common, in the sense that the long run movements will be driven by changes in TFP. Hence, for any vector y_t of endogenous variables, $y_t = Ay_t^x + y_t^c$, where y_t^x is a scalar process, A is a vector of loading and y_t^c is a vector of cycles. If y_t has the representation $\Delta y_t = \bar{y} + D(\ell)e_t$, the (common) trend component of y_t can be obtained using $D(1)e_t = [1,\ldots,1]'e_t^x$, where e_t^x is the innovation in the permanent component.

Exercise 3.44 *Suppose there exists a structural model of the form* $\Delta y_t = \mathcal{A}(\ell)y_{t-1} + \mathcal{A}_0\epsilon_t$, *where* $E(\epsilon_{it}\epsilon'_{i't}) = 0, \quad \forall i, i'$. *Show how to compute* y_t^c.

Exercise 3.45 *Consider a trivariate system including* $(\Delta GDP, c/GDP, inv/GDP)$ *and suppose that* ϵ_{1t} *has long run effects only on* GDP. *Show that if the RBC model is correct* $c/GDP, inv/GDP$ *are stationary. How would you identify a permanent and two transitory shocks?*

BQ and KPSW procedures are similar. However, in the latter, more information is used to estimate the trend, including cointegration restrictions and a larger number of variables. Also, the KPSW approach is easily generalized to large systems while the BQ decompositions is primarily designed for bivariate models.

The KPSW decomposition is also similar to the BN decomposition. The major difference is the "behavioral" content of identified shocks: here the trend is driven by one of the identified shocks of the system, while in the BN decomposition Δy_t^x is driven by a combination of all reduced form shocks.

Example 3.24 *General equilibrium models that extensively exploit BQ and KPSW decompositions to identify permanent "behavioral" shocks are somewhat difficult to construct since multiple permanent shocks may not be separately identifiable. One exception is the two country RBC model of Ahmed et al. (1993). Here output is produced via* $GDP_{it} = K_{it}^{1-\eta}(\zeta_t^{b_1} N_{it})^{\eta}, i = 1, 2$ *where* $\Delta \ln \zeta_t = \bar{\zeta} + \epsilon_{1t}$ *is the common world technology shock and* b_1 *measures the (asymmetric) impact of the shock in the two countries (i.e* $b_1 = 1$ *if* $i = 1$ *and* $b_1 < 1$ *if* $i = 2$*). Labor supply is exogenously given (in the long run) by* $\Delta \ln N_{it} = \bar{N} + \epsilon_{2t}^i$*. Governments consume an exogenously given amount* $g_{it} \equiv \frac{G_{it}}{GDP_{it}} = g_{it-1} + \epsilon_{3t}^i + b_2 \epsilon_{3t}^{i'}$*, where* b_2 *captures the comovements of the shocks in the two countries. Representative agents in country* i *maximize* $E_t \sum_t \beta^t [\upsilon_{it} \ln C_{it} + \upsilon_{i't} \ln C_{i't} + V(N_{it})]$ *where* $\frac{\upsilon_{i't}}{\upsilon_{it}}$ *measures the extent of home bias in consumption. We assume that* $\ln \upsilon_{it}$ *are random walk with disturbances* $\epsilon_{4t}^i, i = 1, 2$*. Finally, the growth rate of relative money supplies evolves according to* $\Delta \ln M_{1t} - \Delta \ln M_{2t} = b_4 + b_6 \epsilon_{1t} + b_5 \epsilon_{2t}^1 + b_7 \epsilon_{2t}^2 + b_8[(1 - b_2)(\epsilon_{3t}^2 - \epsilon_{3t}^1] + b_9(\epsilon_{4t}^2 - \epsilon_{4t}^1) + b_{10}(1 - b_5^1)(\epsilon_{5t}^1 - \epsilon_{5t}^2)$*, where* ϵ_{5t}^i *are money demand shocks.*

Let $p_t = p_{1t}^{b_3} p_{2t}^{1-b_3}$ *and let the relative price of foreign goods in terms of domestic price be* $Tot_t = \frac{p_{2t}}{p_{1t}}$*. The model delivers an expression for the evolution of private output* (GDP_{it}^p) *which can be added to those determining aggregate domestic labor supply, total output in the two countries, relative money supplies and the terms of trade to produce a system of the form* $\Delta y_t - \bar{y} = \mathcal{D}_0 \epsilon_t$ *where* $\Delta y_t = [\Delta \ln N_{1t}, \Delta \ln GDP_{1t}, \Delta \ln GDP_{2t}, \Delta \ln GDP_{1t}^p - \Delta \ln GDP_{1t}^p, \Delta \ln Tot_t, \Delta \ln M_{1t} - \Delta \ln M_{2t}]$*,* $\epsilon_t = [\epsilon_{2t}^1, \epsilon_{1t}, \epsilon_{2t}^2, (1-b_2)(\epsilon_{3t}^2 - \epsilon_{3t}^1), \epsilon_{4t}^2 - \epsilon_{4t}^1, b_{10}(\epsilon_{5t}^1 - \epsilon_{5t}^2]$*;* $\mathcal{D}_0 = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & b_1 & 1 & 0 & 0 & 0 \\ 1 & 1 - b_1 & -1 & 1 & 0 & 0 \\ b_3 & b_3(1 - b_1) & -b_3 & b_3 & b_3 & 0 \\ b_5 & b_6 & b_7 & b_8 & b_9 & 1 \end{bmatrix}$ *and* $\bar{y} = [\bar{N}_1, \bar{\zeta} + \bar{N}_1, b_1 \bar{\zeta} + \bar{N}_2, (\bar{N}_1 - \bar{N}_2) + (1 - b_1)\bar{\zeta}, b_3[(\bar{N}_1 - \bar{N}_2) + (1 - b_1)\bar{\zeta}], b_4]'$*. Given the (restricted) lower triangular structure of* $\mathcal{D}_0$*, it is possible to obtain 6 long run shocks. Methods to identify them are described in chapter 4. Note that, because there is no cointegrating relationship, all shocks have permanent effects on* y_t*. Hence, the permanent-transitory decomposition produced by this model is trivial (*$\Delta y_t^c = 0$ *for all* t*.) Clearly, if a general system* $\Delta y_t = D(\ell) e_t$ *is estimated,* $\Delta y_t^c \neq 0$*. Note that a BN decomposition for this latter system is* $\Delta y_t = D(1) e_t + \frac{(D(\ell) - D(1)}{1 - \ell} \Delta e_t$*, where* e_t *are reduced form shocks.*

3.4 Time Aggregation and Cycles

A problem not fully appreciated in the literature occurs when data is time aggregated. In fact, time series which show important high frequency periodicities may display significant power at business cycle frequencies, when the data is time aggregated. Time aggregation is essentially a two-step filter. In the first step, the variable under consideration is passed through a one-sided filter $\mathcal{B}(\ell) = 1 + \ell_\ell^2 + \ldots + \ell^{n-1}$, if averages over n periods are taken, or $\mathcal{B}(\ell) = \ell^k, k = \{0, 1, , \ldots, n-1\}$ if systematic sampling takes place. In the second step, one typically samples $\mathcal{B}(\ell)y_t$ every n-th observation to obtain non-overlapping aggregates.

In terms of spectra, a time aggregated series is related to its original counterpart via the folding operator $\mathfrak{F}(\mathcal{S}(\omega)) = \sum_{j=-I}^{I} \mathcal{S}(\omega + \frac{2\pi j}{n})$ where $\mathfrak{F}(\mathcal{S}(\omega))$ is defined over $\omega = [-\frac{\pi}{n}, \frac{\pi}{n}]$ and I is the largest integer such that $(\omega + \frac{2\pi j}{n} \in [-\omega, \omega])$. The folding operator reflects the aliasing problem where harmonics of the various frequencies can not be distinguished from one another in the sample data. In essence, aliasing implies that frequencies outside $[-\frac{\pi}{n}, \frac{\pi}{n}]$ in the time aggregated process are folded back inside the $[-\frac{\pi}{n}, \frac{\pi}{n}]$ range. Then $\mathcal{S}_{y^{TA}}(\omega) = \mathfrak{F}(|\mathcal{B}(\ell)|^2 \mathcal{S}(\omega))$.

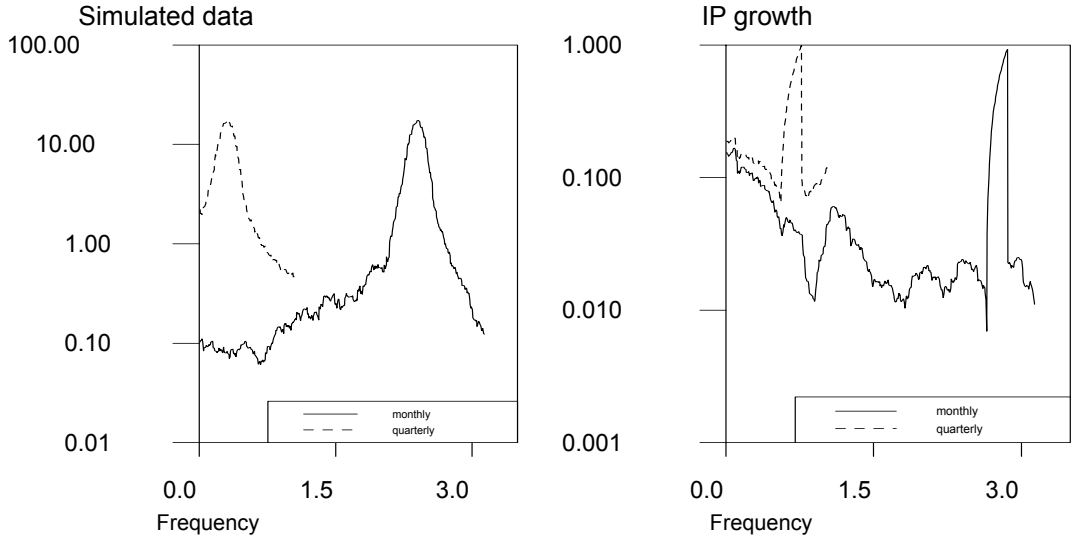


Figure 3.6: Monthly and Quarterly spectrum, simulated data and IP growth

Example 3.25 *Using the folding operator, it is easy to show that a series which shows no power at business cycle frequencies (meaning that no peaks in the spectral density at frequency range corresponding to cycles from 18 to 96 months is visible) with monthly data*

has power at these frequencies when quarterly aggregates are constructed. Consider, for example, the process depicted in the left hand side panel of figure 3.6 which has most of its power in the area around $\omega = 2.2$ (slightly less than 3 months cycles). If we quarterly aggregate this series the area between $\omega = 1.05$, and $\omega = 3.14$ will be folded over the range $\omega \in [0.0, 1.04]$. Hence, the spectrum of the quarterly series has a peak around $\omega = 0.20$, which correspond to cycles of roughly 30 quarters (see dotted line).

Time series which have these features are not unusual. For example, in the right panel of figure 3.6 we plot the spectrum of the US industrial production growth series using monthly and quarterly data. Clearly, the same phenomena occurs.

Exercise 3.46 *Using stock returns for G-7 countries examine whether time aggregation creates spurious peaks at business cycle frequencies using monthly, quarterly and annual data.*

3.5 Collecting Cyclical Information

Once the components of a time series are obtained, statistics summarizing their features can be computed and reported. Typically, two complementary scopes should be balanced. First, statistics should contain sufficient information to allow policymakers and practitioners to assess the state of the economy. Second, they should summarize the characteristics of the cycle efficiently to allow academics to distinguish between different theoretical models of propagation.

While the growth literature has concentrated on great ratios (consumption to output, investment in physical and human capital to output, saving ratios, etc.), the business cycle literature has typically focused attention on the autocovariance function of y_t^c. Hence, in general, variability, auto and cross correlations are presented. When a detailed description of the variability and of the correlations at various frequencies is of interest, spectral densities or bivariate coherence measures are reported. All these statistics can be analytically computed from the ACF of y_t if the form of the cyclical filter $\mathcal{B}^c(\ell)$ is known.

Example 3.26 *Let $\mathcal{B}^c(\ell)$ be a cyclical filter and let $ACF_y(\tau)$ be the autocovariance function of y_t. The autocovariance function of $y_t^c = \mathcal{B}^c(\ell)y_t$ is $ACF_{y^c}(\tau) = ACF_y(0)\sum_{i=-\infty}^{\infty}\mathcal{B}_i^c\mathcal{B}_{i-\tau}^c + \sum_{\tau'=1}^{\infty}ACF_y(\tau')\sum_{i=-\infty}^{\infty}\mathcal{B}_i^c\mathcal{B}_{i-\tau'-\tau}^c + \sum_{\tau'=1}^{\infty}[ACF_y(\tau')]'\sum_{i=-\infty}^{\infty}\mathcal{B}_i^c\mathcal{B}_{i-\tau'-\tau}^c$. The actual computation of $ACF_c(\tau)$ requires truncation of the above expressions, for example, letting i go from $\underline{i}$ to $\bar{\imath}$ and letting τ' go from 1 to $\bar{\tau}'$. For some cyclical filters (e.g. the growth filter) no truncation is needed. In general, a researcher needs to take a stand on the truncation points and, for comparability, it is important to clearly state what $\bar{\tau}'$, $\underline{i}$ and $\bar{\imath}$ are.*

Apart from differences in truncation points, different studies may report different cyclical statistics because the $\mathcal{B}^c(\ell)$ used is different. As we have seen, $\mathcal{B}^c(\ell)$ are different because (a) some decompositions use univariate while others multivariate information (hence, that estimates of y_t^c may have different precision); (b) some decompositions impose orthogonality between the components while others do not (hence, estimates of y_t^c have different spectra);

(c) some decompositions produce estimated long run components with both business cycle as well as high frequency variability (so that the average periodicity of estimated cycles is different); and (d) the weights used in different cyclical filters differ - they could be tent-like or wave-like, they could be symmetric or asymmetric, they may be truncated or exact, etc.. Hence, it is by mere accident that alternative methods produce cyclical components with similar ACFs. While there is a tendency to sweep these differences under the rug or treat them unimportant, major discrepancies may result.

Example 3.27 *We repeat the exercise of Canova (1998) and pass standard US macroeconomic series through a number of filters. In table 3.2 we report the results produced using HP filter with* $\lambda = 1600$ *and* $\lambda = 4$*, the BN filter, a frequency domain BP filter and the KPSW filter using US data from 1955:1 to 1986:4.*

Few features deserve attention. First, not only the variability of each series but also the relative ranking of variabilities changes with the method. Particularly striking is the alteration of the relative variability in the real wage. Second, the magnitude of correlations differs significantly (see e.g. the consumption output correlations). Third, the two HP filters deliver different statistics and the HP4 results mimic those of the BN filter. Finally, the average periodicity of the GDP cycles produced is substantially different.

These difference are also present in other statistics. For example Canova (1999) shows that the dating of cyclical turning points is not robust and that, apart from the BP filter, most methods generate several false alarms relative to the standard NBER classification.

	Variability	Relative Variability		Contemporaneous Correlations			Periodicity
Method	GDP	Consumption	Real wage	(GDP,C)	(GDP,Inv)	(GDP, w)	(quarters)
HP1600	1.76	0.49	0.70	0.75	0.91	0.81	24
HP4	0.55	0.48	0.65	0.31	0.65	0.49	7
BN	0.43	0.75	2.18	0.42	0.45	0.52	5
BP	1.14	0.44	1.16	0.69	0.85	0.81	28
KPSW	4.15	0.71	1.68	0.83	0.30	0.89	6

Table 3.2: Summary statistics

While somewhat disturbing, the results of example 3.27 should not be a deterrent to conscientious researchers and do not support the claim that "there are no business cycle facts" to compare models to. On the contrary, observed differences stress the need to properly define a criteria to assess the empirical relevance of various decompositions. If it were possible to know with reasonable precision what assumptions characterize observables and unobservables, e.g. whether the series is integrated or the trend deterministic; whether the trend is cyclical or not; whether relevant cycles have constant median periodicity or not, one could select among methods on the basis of some statistical optimality criteria (e.g. minimization of the MSE). Given that this is not possible the literature has arbitrarily concentrated on an economic criteria: interesting periodicities are those within 6 and 32 quarters. However, even with this focus, care should be exercised for three reasons. First,

within the class of methods which extract cycles with these periodicities, differences may emerge if variables have most of their spectral power concentrated in a neighborhood of the trend/cycle cut-off point (see Canova (1998) for an example involving the hours and productivity correlation). Second, especially in international comparisons, results may differ if a series has different cyclical periodicity across countries. Third, filters which allow to single out these frequencies, may produce spurious periodicity and correlations.

Exercise 3.47 *Simulate data for the output gap, inflation and the nominal interest rate from a basic sticky price model (e.g. the one whose solution is given in exercise 2.31, after you proxy marginal costs with the output gap and use the identity that consumption equal to the output gap) driven by three shocks: a monetary policy, a cost push and an Euler equation shock. Choose the parameters appropriately and compute the spectral density of simulated data and for actual EU data. How do they compare? Repeat the exercise using HP and Band Pass filtered versions of the actual data.*

These arguments bring us to an important aspect of the comparison between the actual data and the data produced by DSGE models. It is often stressed that, for comparability, the same filtering approach should be used to compute the ACF of y_t^c in both data. Rarely, however, this principle takes into account the (known) properties of simulated data. For example, in several of the models of chapter 2, the simulated series inherit the properties of the driving forces (they are persistent if shocks are; integrated if shocks are, etc.). If the purpose of filtering is the removal of the trend, no transformation (or a growth transformation) should then be applied to simulated data. On the other hand, if the purpose of filtering is to bring out cycles with certain periodicity, one should rememeber that approximate BP or HP filters may produce distortions in highly persistent series like the one typically produced by DSGE models. Hence, it is not very difficult to build examples where simulated and actual filtered data look similar at business cycle frequencies even though the model and the data are substantially at odds with each other.

Example 3.28 *Consider two AR(1) processes:* $y_{1t} = 1.0 + 0.8y_{1t-1} + e_t,\ e_t \sim N(0,1)$ *and* $y_{2t} = 1.0 - 0.55y_{2t-1} + e_t,\ e_t \sim N(0,1)$ *The spectra of the two series are specular but if the shocks to the two series are the same, their variability at business cycle frequencies will be similar. In fact, BP filtered variabilities are 1.32 in both cases. Clearly, the two DGPs have very different features.*

Hence, it is not clear that comparability is a relevant criterion and, for some purposes, it may be more reasonable to filter actual data but not the simulated one or filter the two types of data with different filters, especially if the model is not assumed to be the correct process generating the actual data.

All in all, mechanical application of filters is dangerous. If one insists on trying to compare actual and simulated data using second moments, one should carefully look into the properties of the data and be aware of the features of the cycle extracting filter used.

The alternative is to shift attention away from the second moments of the growth cycle and as in Pagan and Harding (2002), (2005), use statistics which can be obtained directly from the observable series (see also Hess and Iwata (1997) and King and Plosser (1989)).

The approach is closely related to Burns and Mitchell (1946) methodology and requires the identification of turning points in the "reference" variable (say, GDP or an aggregate of important macro series), the measurement of durations, amplitudes and cumulative changes of the cycles and of their phases, and the documentation of asymmetries over various phases. All these statistics can be computed from the (log) level of y_t using a version of the so-called Bry and Boschen (1971) algorithm, which we describe next:

Algorithm 3.4

1) *Smooth y_t with a series of filters (to eliminate outliers, high frequency variations, irregular or uninteresting fluctuations). Call y_t^{sm} the smoothed series.*

2) *Use a dating rule to determine a potential set of turning points. One simple rule is* $\Delta^2 y_t^{sm} > 0 (< 0), \Delta y_t^{sm} > 0 (< 0), \Delta y_{t+1}^{sm} < 0 (> 0), \Delta^2 y_{t+1}^{sm} < 0 (> 0)$.

3) *Use a censoring rule to ensure that peaks and throughs alternate and that the duration and the amplitude of phases is meaningful.*

Hence, to obtain turning points and business cycle phases, one needs to make some choices. While there are differences in the literature, the consensus is that a two-quarter rule like the one used in 2) or slight variations of it (see e.g. Lahiri and Moore (1991)) suffices to date turning points. As for censoring rules, it is typical to impose a minimum duration of each phase of two or three quarters, so that complete cycles should be at least 5 to 7 quarters long, and/or some minimum amplitude restriction, e.g. peaks to throughs drops of less than one percent should be excluded. Note also that the first step could be dispensed of if the censoring rule in 3) is strong enough.

Once turning points are identified one could compute a number of interesting statistics. For example, average durations (AD), i.e. average length of time spent between peaks or between peaks and throughs, average amplitudes (AA), i.e. the average size of the drop between peaks and throughs or of the gain between troughs and peaks; average cumulative changes over phases ($CM = 0.5 * (AD * AA)$) and excess average cumulative changes ($(CM - CM^A + 0.5 * AA)/AD$), where CM^A is the actual average cumulative change. Finally, one can compute a concordance index $CI_{i,i'} = n^{-1}[\sum \mathcal{I}_{i't}\mathcal{I}_{it} - (1 - \mathcal{I}_{i't})(1 - \mathcal{I}_{it})]$, which can be used to assess the strenght of the comovements of two variables over business cycle phases. Here n is the number of complete cycles, $\mathcal{I}_{it} = 1$ in expansions and $\mathcal{I}_{it} = 0$ in contractions. Note that $CI_{i,i'} = 1$ if the two series are in the same phase at all times and zero if the series are perfectly negatively correlated.

Example 3.29 *Applying the dating rule described in algorithm 3.4 with a minimum duration of the cycle of 5 quarters, to US output, US consumption and US investment for the*

Variable	*Duration (quarters)*		*Amplitude (percentage)*		*Excess change (percentage)*		*Concordance (percentage)*
	PT	*TP*	*PT*	*TP*	*PT*	*TP*	
GDP	*3*	*18.7*	*-2.5*	*20.7*	*-0.1*	*1.1*	
C	*2.9*	*38*	*-2.0*	*39*	*0.2*	*0.1*	*0.89*
Inv	*5.2*	*11.1*	*23.3*	*34.7*	*1.7*	*2.7*	*0.78*

Table 3.3: US Business Cycle Statistics

period 1947:1-2003:1 the statistics contained in table 3.3 are obtained. Expansion phases are under the heading TP and recession phases under the heading PT.

On average, expansions in consumption are much longer and much stronger than those in GDP. Also, investment displays a much stronger change in expansions than GDP but shorter average duration and relatively long contraction phases. In general, asymmetries over cyclical phases are present in all three series.

Exercise 3.48 *Repeat the calculations performed in example 3.29 using i) the dating rule* $\Delta y_t^{sm} > 0, \Delta y_{t+1}^{sm} < 0$ *to find peaks and* $\Delta y_t^{sm} < 0, \Delta y_{t+1}^{sm} > 0$ *to find throughs; ii) requiring a minimum duration of 7 quarters for the full cycle.*

Exercise 3.49 *Using Euro area data for output, consumption and investment for the period 1970:1-2002:4 calculate the same four statistics presented in example 3.29 and compare them to the one of the US. Is there any interesting pattern that makes the Euro area different?*

There are two appealing features of the approach. First, cyclical statistics can be obtained without extracting cyclical components. Second, they can be computed even when no cycle exists, in the sense that all shocks have permanent effects or that y_t does not have any power at business cycle frequencies. This latter feature is important in comparing DSGE models and the data. In fact, barring few cases, the models described in chapter 2 produce (approximate) VAR(1) solutions. Hence, the data produced by the model does not display peaks in the spectral density at cyclical frequencies, and therefore is ill-suited to be compared to the data using decompositions which look for important periodicities or simply emphasize business cycle frequencies. A couple of drawbacks should also be mentioned. First, statistics may be sensitive to dating and censoring rules. Since the rules of algorithm 3.4 are arbitrary, one should carefully monitor the sensitivity of turning point dates to the choices made. Second, it is not clear how to adapt dating and censoring rules when international comparisons needs to be made.

Finally, one should remember that both second moments and turning point statistics provide reduced form information. That is, they are uninformative about comovements in response to economically interesting shocks and silent about the sources of cyclical fluctuations. This kind of conditional information is exactly what structural VARs, considered in the next chapter, deliver.

Chapter 4: VAR Models

This chapter describes a set of techniques which stand apart from those considered in the next three chapters, in the sense that economic theory is only minimally used in the inferential process. VAR models, pioneered by Chris Sims about 25 years ago, have acquired a permanent place in the toolkit of applied macroeconomists both to summarize the information contained in the data and to conduct certain types of policy experiments. VAR are well suited for the first purpose: the Wold theorem insures that any vector of time series has a VAR representation under mild regularity conditions and this makes them the natural starting point for empirical analyses. We discuss the Wold theorem, and the issues connected with non-uniqueness, non-fundamentalness and non-orthogonality of the innovation vector in the first section. The Wold theorem is generic but imposes important restrictions; for example, the lag length of the model should go to infinity for the approximation to be "good". Section 2 deals with specification issues, describes methods to verify some of the restrictions imposed by the Wold theorem and to test other related implications (e.g. white noise residuals, linearity, stability, etc.). Section 3 presents alternative formulations of a VAR(q). These are useful when computing moments or spectral densities, and in deriving estimators for the parameters and for the covariance matrix of the shocks. Section 4 presents statistics commonly used to summarize the informational content of VARs and methods to compute their standard errors. Here we also discuss generalized impulse response functions, which are useful in dealing with time varying coefficients VAR models analyzed in chapter 10. Section 5 deals with identification, i.e with the process of transforming the information content of reduced form dynamics into structural ones. Up to this point economic theory has played no role. However, to give a structural interpretation to the estimated relationships, economic theory needs to be used. Contrary to what we will be doing in the next three chapters, only a minimalist set of restrictions, loosely related to the classes of models presented in chapter 2, are employed to obtain structural relationships. We describe identification methods which rely on conventional short run, on long run and on a sign restrictions. In the latter two cases (weak) restrictions derived from DSGE models are employed and the structural link between the theory and the data explicitly made. Section 6 describes problems which may distort the interpretation of structural VAR results. Time aggregation, omission of variables and shocks and non-fundamentalness should always be in the back of the mind of applied researchers when conducting policy analyses with VAR. Section 7 proposes a way to validate a class of DSGE models using structural

VARs. Log-linearized DSGE models have a restricted VAR representation. When a researcher is confident in the theory, a set of quantitative restrictions can be considered, in which case the methods described in chapters 5 to 7 could be used. When theory only provides qualitative implications or when its exact details are doubtful, one can still validate a model conditioning on its qualitative implications. Since DSGE models provide a wealth of robust sign restrictions, one can take the ideas of section 5 one step further, and use them to identify structural shocks. Model evaluation then consists in examining the qualitative (and quantitative) features of the dynamic responses to identified structural shocks. In this sense, VAR identified with sign restrictions offer a natural setting to validate incompletely specified (and possibly false) DSGE models.

4.1 The Wold theorem

The use of VAR models can be justified in many ways. Here we employ the Wold representation theorem as major building block. While the theory of Hilbert spaces is needed to make the arguments sound, we keep the presentation simple and invite the reader to consult Rozanov (1967) or Brockwell and Davis (1991) for precise statements.

The Wold theorem decomposes any $m\times 1$ vector stochastic process $y_t^\dagger$ into two orthogonal components: one linearly predictable and one linearly unpredictable (linearly regular). To show what the theorem involves let $\mathcal{F}_t$ be the time t information set; $\mathcal{F}_t = \mathcal{F}_{t-1} \oplus \mathcal{E}_t$, where $\mathcal{F}_{t-1}$ contains time $t-1$ information and $\mathcal{E}_t$ the news at t. Here $\mathcal{E}_t$ is orthogonal to $\mathcal{F}_{t-1}$ (written $\mathcal{E}_t \perp \mathcal{F}_{t-1}$) and $\oplus$ indicates direct sum, that is $\mathcal{F}_t = \{y_{t-1}^\dagger + e_t, y_{t-1}^\dagger \in \mathcal{F}_{t-1},\ e_t \in \mathcal{E}_t\}$.

Exercise 4.1 *Show that $\mathcal{E}_t \perp \mathcal{F}_{t-1}$ implies $\mathcal{E}_t \perp \mathcal{E}_{t-1}$ so that $\mathcal{E}_{t-j}$ is orthogonal to $\mathcal{E}_{t-j'}$, $j' < j$.*

Since the decomposition of $\mathcal{F}_t$ can be repeated for each t, iterating backwards we have

$$\mathcal{F}_t = \mathcal{F}_{t-1} \oplus \mathcal{E}_t = \ldots = \mathcal{F}_{-\infty} \oplus \sum_{j=0}^{\infty} \mathcal{E}_{t-j} \tag{4.1}$$

where $\mathcal{F}_{-\infty} = \bigcap_j \mathcal{F}_{t-j}$. Since $y_t^\dagger$ is known at time t (this condition is sometimes referred as adaptability of $y_t^\dagger$ to $\mathcal{F}_t$), we can write $y_t^\dagger \equiv E[y_t^\dagger|\mathcal{F}_t]$ where $E[.|\mathcal{F}_t]$ is the conditional expectations operator. Orthogonality of the news with past information then implies:

$$y_t^\dagger = E[y_t^\dagger|\mathcal{F}_t] = E[y_t^\dagger|\mathcal{F}_{-\infty} \oplus \sum_j \mathcal{E}_{t-j}] = E[y_t^\dagger|\mathcal{F}_{-\infty}] + \sum_{j=0}^{\infty} E[y_t^\dagger|\mathcal{E}_{t-j}] \tag{4.2}$$

We make two assumptions. First, we consider linear representations, that is, we substitute the expectations operator with a linear projection operator. Then (4.2) becomes

$$y_t^\dagger = a_t y_{-\infty} + \sum_{j=0}^{\infty} D_{jt} e_{t-j} \tag{4.3}$$

where $e_{t-j} \in \mathcal{E}_{t-j}$ and $y_{-\infty} \in \mathcal{F}_{-\infty}$. The sequence $\{e_t\}_{t=0}^{\infty}$, defined by $e_t = y_t^\dagger - E[y_t^\dagger|\mathcal{F}_{t-1}]$, is a white noise process (i.e. $E(e_t) = 0$; $E(e_t e_{t-j}') = \Sigma_t$ if $j = 0$ and zero otherwise). Second, we assume that $a_t = a$; $D_{jt} = D_j$; $\forall t$. This implies

$$y_t^\dagger = a y_{-\infty} + \sum_{j=0}^{\infty} D_j e_{t-j} \tag{4.4}$$

Exercise 4.2 *Show that if $y_t^\dagger$ is covariance stationary, $a_t = a$, $D_{jt} = D_j$.*

The term $ay_{-\infty}$ on the right hand side of (4.4) is the linearly deterministic component of $y_t^\dagger$ and can be perfectly predicted given the infinite past. The term $\sum_j D_j e_{t-j}$ is the linearly regular component, that is, the component produced by the news at each t. We say that $y_t^\dagger$ is deterministic if and only if $y_t^\dagger \in \mathcal{F}_{-\infty}$ and regular if and only if $\mathcal{F}_{-\infty} = \{0\}$.

Three important points need to be highlighted. First, for (4.2) to hold, no assumptions about $y_t^\dagger$ are required: we only need that new information is orthogonal to the existing one. Second, both linearity and stationary are unnecessary for the theorem to hold. For example, if stationarity is not assumed there will still be a linearly regular and a linearly deterministic component even though each will have time varying coefficients (see (4.3)). Third, if we insist on requiring covariance stationary, preliminary transformations of $y_t^\dagger$ may be needed to produce the representation (4.4).

The Wold theorem is a powerful tool but is too generic to guide empirical analysis. To impose some more structure, we assume first that the data is a mean zero process, possibly after deseasonalization (with deterministic periodic functions), removal of constants, etc. and let $y_t = y^\dagger - a y_{-\infty}$. Using the lag operator we write $\sum_{j=0}^{\infty} D_j e_{t-j} = \sum_j D_j \ell^j e_t = D(\ell) e_t$ so that $y_t = D(\ell) e_t$ is the MA representation for y_t where D_j is a $m \times m$ matrix of rank m, for each j. MA representations are not unique: in fact, for any nonsingular matrix $\mathcal{H}(\ell)$ satisfying $\mathcal{H}(\ell)\mathcal{H}(\ell^{-1})' = I$ such that $\mathcal{H}(z)$ has no singularities for $|z| \leq 1$, where $\mathcal{H}(\ell^{-1})'$ is the transpose (and possibly complex conjugate) of $\mathcal{H}(\ell)$, we can write $y_t = \tilde{D}(\ell)\tilde{e}_t$ with $\tilde{D}(\ell) = D(\ell)\mathcal{H}(\ell)$, $\tilde{e}_t = \mathcal{H}(\ell^{-1})' e_t$.

Exercise 4.3 *Show that $E(\tilde{e}_t \tilde{e}_{t-j}') = E(e_t e_{t-j}')$. Conclude that if e_t is covariance stationary, the two representation produce equivalent autocovariance functions for y_t.*

Matrices like $\mathcal{H}(\ell)$ are called Blaschke factors and are of the form $\mathcal{H}(\ell) = \prod_{i=1}^{m} \varrho_i \mathcal{H}^\dagger(d_i, \ell)$ where d_i are the roots of $D(\ell)$, $|d_i| < 1$, $\varrho_i \varrho_i' = I$ and, for each i, $\mathcal{H}^\dagger(d_i, \ell)$ is given by:

$$\mathcal{H}^\dagger(d_i, \ell) = \begin{bmatrix} 1 & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & \frac{\ell - d_i}{1 - d_i^{-1}\ell} & \dots & 0 \\ 0 & 0 & \dots & 1 \end{bmatrix} \tag{4.5}$$

Exercise 4.4 *Suppose* $\begin{pmatrix} y_{1t} \\ y_{2t} \end{pmatrix} = \begin{pmatrix} (1+4\ell) & 0 \\ 0 & (1+10\ell) \end{pmatrix} \begin{pmatrix} y_{1t} \\ y_{2t} \end{pmatrix}$. *Find the Blaashke factors of $D(\ell)$. Construct two alternative moving average representations for y_t.*

Example 4.1 *Consider $y_{1t} = e_t - 0.5e_{t-1}$ and $y_{2t} = \tilde{e}_t - 2\tilde{e}_{t-1}$. It is easy to verify that the roots of $D(z)$ are 2 in the first case, and 0.5 in the second. Since the roots are one the inverse of the other, the two processes span the same information space as long as the variance of innovations is appropriately adjusted. In fact, using the covariance generating function to have $CGF_{y_1}(z) = (1 - 0.5z)(1 - 0.5z^{-1})\sigma_1^2$ and $CGF_{y_2}(z) = (1 - 2z)(1 - 2z^{-1})\sigma_2^2 = (1 - 0.5z)(1 - 0.5z^{-1})(4\sigma_2^2)$. Hence, if $\sigma_1^2 = 4\sigma_2^2$ the CGF of the two processes is the same.*

Exercise 4.5 *Let $y_{1t} = e_t - 4e_{t-1},\ e_t \sim (0, \sigma^2)$. Set $y_{2t} = (1 - 0.25\ell)^{-1}y_{1t}$. Show that the CGF(z) of y_{2t} is a constant for all z. Show that $y_{2t} = \tilde{e}_t - 0.25\tilde{e}_{t-1}$ where $\tilde{e}_t \sim (0, 16\sigma^2)$ is equivalent to y_{1t} in terms of the covariance generating function.*

Among the class of equivalent MA representations, it is typical to choose the "fundamental" one. The following two definitions are equivalent.

Definition 4.1 *(Fundamentalness)*
1) A MA is fundamental if $det(D_0E(e_te_t')D_0') > det(D_jE(e_{t-j}e_{t-j}')D_j'),\ \forall j \neq 0$.
2) A MA is fundamental if the roots of $D(z)$ are all greater than one in modulus.

The roots of $D(z)$ are related to the eigenvalues of the companion matrix of the system (see section 3). Fundamental representations, also termed Wold representations, could also be identified by the requirement that the completion of the space spanned by linear combinations of the y_t's has the same information as the completion of the space spanned by linear combinations of e_t's. In this sense Wold representations are invertible: knowing y_t is the same as knowing e_t.

As it is shown in the next example, construction of a fundamental representation requires "flipping" all roots that are less than one in absolute value.

Example 4.2 *Suppose $y_t = \begin{bmatrix} 1.0 & 0 \\ 0.2 & 0.9 \end{bmatrix} e_t + \begin{bmatrix} 2.0 & 0 \\ 0 & 0.7 \end{bmatrix} e_{t-1}$ where $e_t \sim$ iid $(0, I)$. Here $det(D_0) = 0.9 < det(D_1) = 1.4$ so the representation is not fundamental. To find a fundamental one we compute the roots of $D_0 + D_1z = 0$; their absolute values are 0.5 and 1.26 (these are the diagonal elements of $-D_1^{-1}D_0$). The problematic root is 0.5 which we flip to 1.0/0.5=2.0. The fundamental MA is then $y_t = \begin{bmatrix} 1.0 & 0 \\ 0.2 & 0.9 \end{bmatrix} e_t + \begin{bmatrix} 0.5 & 0 \\ 0 & 0.7 \end{bmatrix} e_{t-1}$.*

Exercise 4.6 *Determine which of the following polynomial produces fundamental representations when applied to a white noise innovation:(i) $D(\ell) = 1 + 2\ell + 3\ell^2 + 4\ell^3$, (ii) $D(\ell) = 1 + 2\ell + 3\ell^2 + 2\ell^3 + \ell^4$, (iii) $D(\ell) = I + \begin{bmatrix} .8 & -.7 \\ .7 & .8 \end{bmatrix} \ell$, (iv) $D(\ell) = \begin{bmatrix} 1 & 1 \\ 3 & 4 \end{bmatrix} + \begin{bmatrix} 3 & 2 \\ 4 & 1 \end{bmatrix} \ell + \begin{bmatrix} 4 & 3 \\ 2 & 1 \end{bmatrix} \ell^2$.*

Exercise 4.7 *Show that* $y_t = e_t + \begin{bmatrix} 1.0 & 0 \\ 0 & 0.8 \end{bmatrix} e_{t-1}$ *where* $var(e_t) = \begin{bmatrix} 2.0 & 1.0 \\ 1.0 & 1.0 \end{bmatrix}$ *and* $y_t = e_t + \begin{bmatrix} 0.9091 & 0.1909 \\ 0 & 0.8 \end{bmatrix} e_{t-1}$ *where* $var(e_t) = \begin{bmatrix} 2.21 & 1.0 \\ 1.0 & 1.0 \end{bmatrix}$ *generate the same ACF for* y_t*. Which representation is fundamental?*

Exercise 4.8 *Let* $\begin{pmatrix} y_{1t} \\ y_{2t} \end{pmatrix} = \begin{pmatrix} (1+4\ell) & 1+0.5\ell \\ 0 & (1+5\ell) \end{pmatrix} \begin{pmatrix} e_{1t} \\ e_{2t} \end{pmatrix}$ *where* $e_t = (e_{1t}, e_{2t})$ *has unitary variance. Is the space spanned by linear combinations of the* y_t *and* e_t *the same? If the MA is not fundamental, find a fundamental one.*

While it is typical to use Wold representations in applied work, there are economic models that do not generate a fundamental format. Two are presented in the next examples.

Example 4.3 *Consider a RBC model where households maximize* $E_0 \sum_t \beta^t (\ln(c_t) - \vartheta_N N_t)$ *subject to* $c_t + inv_t \leq GDP_t$; $K_{t+1} = (1-\delta)K_t + inv_t$; $c_t \geq 0$; $inv_t \geq 0; 0 \leq N_t \leq 1$ *where* $0 < \beta < 1$ *and* δ, ϑ_n *are parameters and assume that the production function is* $GDP_t = k_t^{1-\eta} N_t^{\eta} \zeta_t$ *where* $\ln \zeta_t = \ln \zeta_{t-1} + 0.1\epsilon_{1t} + 0.2\epsilon_{1t-1} + 0.4\epsilon_{1t-2} + 0.2\epsilon_{1t-3} + 0.1\epsilon_{1t-4}$*. Such a diffusion of technological innovations is appropriate when e.g., only the most advanced sector employs the new technology (say, a new computer chips) and it takes some time for the innovation to spread to the economy. If* $\epsilon_{1t} = 1$, $\epsilon_{1t+\tau} = 0, \forall \tau \neq 0$ ζ_t *looks like in figure 1. Clearly, a process with this shape does not satisfy the restrictions given in definition 4.1.*

Example 4.4 *Consider a model where fiscal shocks drive economic fluctuations. Typically, fiscal policy changes take time to have effects: between the programming, the legislation and the implementation of, say, a change in income tax rates several months may elapse. If agents are rational they may react to tax changes before the policy is implemented and, conversely, no behavioral changes may be visible when the changes actually take place. Since the information contained in tax changes may have a different timing than the information contained, say, in the income process, fiscal shocks may produce non-Wold representations.*

Whenever economic theory requires non-fundamental MAs, one could use Blaaske factors to flip the representations provided by standard packages, as e.g. in Lippi and Reichlin (1994). In what follows we will consider only fundamental structures and take $y_t = D(\ell)e_t$ be such a representation.

The "innovations" e_t play an important role in VAR analyses. Since $E(e_t|\mathcal{F}_{t-1}) = 0$ and $E(e_t e_t'|\mathcal{F}_{t-1}) = \Sigma_e$, e_t are serially uncorrelated but contemporaneously correlated. This means that we cannot attach a "name" to the disturbances. To do so we need an orthogonal representation for the innovations. Let Σ_e be the covariance matrix of e_t, let $\Sigma_e = \mathcal{P}\mathcal{V}\mathcal{P}' = \tilde{\mathcal{P}}\tilde{\mathcal{P}}'$ where $\mathcal{V}$ is a diagonal matrix and $\tilde{\mathcal{P}} = \mathcal{P}\mathcal{V}^{0.5}$. Then $y_t = D(\ell)e_t$ is equivalent to

$$y_t = \tilde{D}(\ell)\tilde{e}_t \tag{4.6}$$

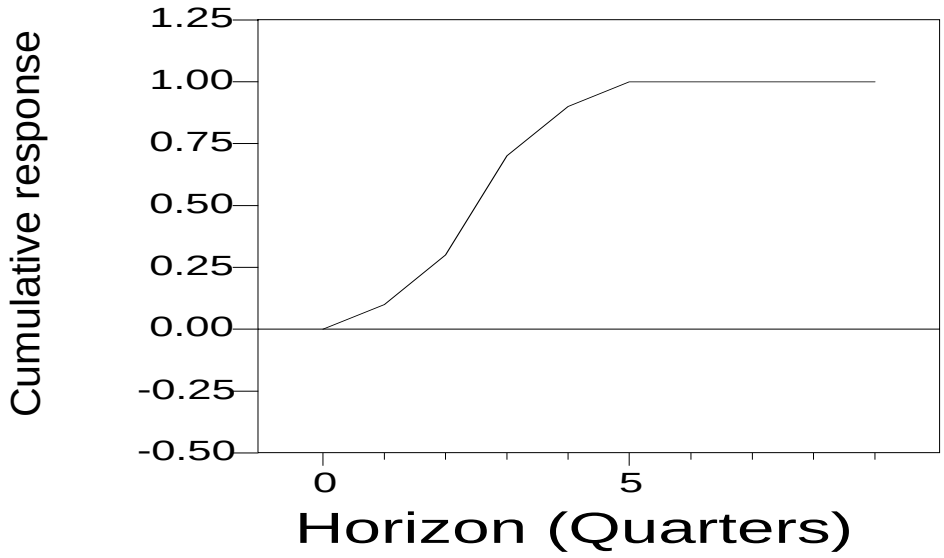


Figure 4.1: Non fundamental technological progress

for $\tilde{D}(\ell) = D(\ell)\tilde{\mathcal{P}}$ and $\tilde{e}_t = \tilde{\mathcal{P}}^{-1}e_t$. There are many ways of generating (4.6). One is a Choleski factorization, i.e. $\mathcal{V} = I$ and $\mathcal{P}$ is a lower triangular matrix. Another is obtained when $\mathcal{P}$ contains the eigenvectors and $\mathcal{V}$ the eigenvalues of Σ_e.

Example 4.5 *If e_t is a 2×1 vector with correlated entries, orthogonal innovations are $\tilde{e}_{1t} = e_{1t} - be_{2t}$ and $\tilde{e}_{2t} = e_{2t}$ where $b = \frac{cov(e_{1t}e_{2t})}{var(e_{2t})}$ and $var(\tilde{e}_{1t}) = \sigma_1^2 - b^2\sigma_2^2$, $var(\tilde{e}_{2t}) = \sigma_2^2$.*

It is important to stress that orthogonalization devices are void of economic content: they only transform the MA representation in a form which is more useful when tracing out the effect of a particular shock. To attach economic interpretations to the representation, these orthogonalizations ought to be linked to economic theory. Note also that while with the Choleski decomposition $\mathcal{P}$ has zero restrictions placed on the upper triangular part, no such restrictions are present when an eigenvalue-eigenvector decomposition is performed.

As mentioned, when the polynomial $D(z)$ has all its roots greater than one in modulus (and this condition holds if, e.g., $\sum_{j=0}^{\infty} D_j^2 < \infty$ (see Rozanov (1967)) the MA representation is invertible and we can express e_t as a linear combination of current and past y_t's, i.e. $[A_0 - A(\ell)]y_t = e_t$ where $[A_0 - A(\ell)] = (D(\ell))^{-1}$. Moving lagged y_t's on the right hand side and setting $A_0 = I$ a vector autoregressive (VAR) representation is obtained

$$y_t = A(\ell)y_{t-1} + e_t \tag{4.7}$$

In general, $A(\ell)$ will be of infinite length for any reasonable specification of $D(\ell)$.

There is an important relationship between the concept of invertibility and the one of stability of the system which we highlight next.

Definition 4.2 *(Stability) A VAR(1) is stable if* $det(I_m - Az) \neq 0, \forall |z| \leq 1$ *and a VAR(q) is stable if* $det(I_m - A_1 z - \ldots - A_q z^q) \neq 0 \ \forall |z| \leq 1$.

Definition 4.2 implies that all eigenvalues of A have modulus less or equal than 1 (or that the matrix A has no roots inside or on the complex unit circle). Hence, if y_t has an invertible MA representation, it also has a stable VAR structure. Therefore, one could start from stable processes to motivate VAR analyses (as, e.g. it is done in Lutkepohl (1991)). Our derivation shows the primitive restrictions needed to obtain stable VARs.

Example 4.6 *Suppose* $y_t = \begin{bmatrix} 0.5 & 0.1 \\ 0.0 & 0.2 \end{bmatrix} y_{t-1} + e_t$. *Here* $det(I_2 - Az) = (1 - 0.5z)(1 - 0.2z) = 0$ *and* $|z_1| = 2 > 1, |z_2| = 5 > 1$. *Hence, the system is stable.*

Exercise 4.9 *Check if* $y_t = \begin{bmatrix} 0.6 & 0.4 \\ 0.5 & 0.2 \end{bmatrix} y_{t-1} + \begin{bmatrix} 0.1 & 0.3 \\ 0.2 & 0.6 \end{bmatrix} y_{t-2} + e_t$ *is stable or not.*

To summarize, any vector of time series can be represented with a constant coefficient VAR(∞) under linearity, stationarity and invertibility. Hence, one can interchangeably think of data or the VAR for the data. Also, with a finite stretch of data only a VAR(q), q finite, can be used. For a VAR(q) to approximate any y_t sufficiently well, we need D_j to converge to zero repidly as j increases.

Exercise 4.10 *Consider* $y_t = e_t + 0.9e_{t-1}$ *and* $y_t = e_t + 0.3e_{t-1}$. *Compute the AR representations. What lag length is needed to approximate the two processes? What if* $y_t = e_t + e_{t-1}$*?*

Two concepts which are of some use are in applied work are those of Granger non-causality and Sims (econometric) exogeneity. It is important to stress that they refer to the ability of one variable to predict another one and do not imply any sort of economic causality (e.g. the government takes an action, the exchange rate will move). Let (y_{1t}, y_{2t}) be a partition of a covariance stationary y_t with fundamental innovations e_{1t} and e_{2t}; let Σ_e be diagonal and let $D_{i,i'}(\ell)$ be the i, i' block of $D(\ell)$.

Definition 4.3 *(Granger causality)* y_{2t} *fails to Granger cause* y_{1t} *if and only if* $D_{12}(\ell) = 0$.

Definition 4.4 *(Sims Exogeneity) We can write* $y_{2t} = \mathsf{Q}(\ell) y_{1t} + \epsilon_{2t}$ *with* $E_t[\epsilon_{2t} y_{1t-\tau}] = 0, \ \forall \tau \geq 0$ *and* $\mathsf{Q}(\ell) = \mathsf{Q}_0 + \mathsf{Q}_1 \ell + \ldots$ *if and only if* y_{2t} *fails to Granger cause* y_{1t} *and* $D_{21}(\ell) \neq 0$.

Exercise 4.11 *Show what Granger non-causality of* y_{2t} *for* y_{1t} *implies in a trivariate VAR.*

We conclude examining cases where the data deviates from the setup considered so far.

Exercise 4.12 *(i) Suppose that $y_t = D(\ell)e_t$ where $D(\ell) = (1-\ell)D^\dagger(\ell)$. Derive a VAR for y_t. Show that if $D^\dagger(\ell) = 1$, there is no convergent VAR representation for y_t.*
(ii) Suppose that $y_t^\dagger = a_0 + a_1 t + D(\ell)e_t$ if $t \leq \bar{T}$ and $y_t^\dagger = a_0 + a_2 t + D(\ell)e_t$ if $t > \bar{T}$. How would you derive a VAR representation for y_t?
(iii) Suppose that $y_t = D(\ell)e_t$ and $var(e_t) \propto y_{t-1}^2$ Find a VAR for y_t.
(iv) Suppose that $y_t = D(\ell)e_t$, $var(e_t) = b\ var(e_{t-1}) + \sigma^2$. Find a VAR for y_t.

4.2 Specification

In section 4.1 we showed that a constant coefficient VAR is a good approximation to any vector of time series. Here we examine how to verify the restrictions needed for the approximation to hold. The model we consider is (4.7) where $A(\ell) = A_1\ell + \ldots + A_q\ell^q$, y_t is a $m \times 1$ vector, and $e_t \sim (0, \Sigma_e)$. VARs with econometrically exogenous variables can be obtained via restrictions on $A(\ell)$ as indicated in definition 4.4. We let $\mathsf{A}_1' = (A_1', \ldots, A_q')'$ be a $(mq \times m)$ matrix and set $\alpha = vec(\mathsf{A}_1)$ where $vec(\mathsf{A}_1)$ stacks the columns of A_1 (so α is a $m^2q \times 1$ vector).

4.2.1 Lag Length 1

There are several methods to select the lag length of a VAR. The simplest is based on a likelihood ratio (LR) test. . Here the model with a smaller number of lags is treated as a restricted version of a larger dimensional model. Since the two models are nested, under the null that the restricted model is correct, differences in the likelihoods should be small. Let $R(\alpha) = 0$ be a set of restrictions and $\mathcal{L}(\alpha, \Sigma_e)$ the likelihood function. Then:

$$\begin{aligned} LR &= 2[\ln \mathcal{L}(\alpha^{un}, \Sigma_e^{un}) - \ln \mathcal{L}(\alpha^{re}, \Sigma_e^{re})] && (4.8) \\ &= (R(\alpha^{un}))'[\frac{\partial R}{\partial \alpha^{un}}(\Sigma_e^{re} \otimes (X'X)^{-1})(\frac{\partial R}{\partial \alpha^{un}})']^{-1}(R(\alpha^{un})) && (4.9) \\ &= T(\ln|\Sigma_e^{re}| - \ln|\Sigma_e^{un}|) \xrightarrow{D} \chi^2(\nu) && (4.10) \end{aligned}$$

where $X_t = (y_{t-1}', \ldots, y_{t-q}')'$, and $X' = (X_0, \ldots, X_{T-1})$ is a $mq \times T$ matrix and ν the number of restrictions. (4.8)-(4.9)-(4.10) are equivalent formulations of the likelihood ratio test. The first is the standard one. (4.9) is obtained maximizing the likelihood function with respect to α subject to $R(\alpha) = 0$. (4.10) is convenient for computing actual test values and to compare LR results with those of other testing procedures.

Exercise 4.13 *Derive (4.9) using a Lagrangian multiplier approach.*

Four important features of LR tests need to be highlighted. First, a LR test is valid when y_t is stationary and ergodic and if the residuals are white noise under the null. Second, it can be computed without explicit distributional assumptions on the y_t's. What is required is that e_t is a sequence of independent white noises with bounded fourth moments and that T is sufficiently large - in which case $\alpha^{un}, \Sigma_e^{un}, \alpha^{re}, \Sigma_e^{re}$ are pseudo maximum likelihood

estimators. Third, a likelihood ratio test is biased against the null in small samples. Hence it is common to use $LR^c = (T-qm)(\ln|\Sigma^{re}| - \ln|\Sigma^{un}|)$ where qm is the number of estimated parameters in each equation of the unrestricted system. Finally, one should remember that the distribution of the LR test is only asymptotically valid. That is, significance levels only approximate probabilities of Type I errors.

In practice, an estimate of q is obtained sequentially as the next algorithm shows:

Algorithm 4.1

1) *Choose an upper bound* $\bar{q}$.

2) *Test a VAR*$(\bar{q}-1)$ *against VAR*$(\bar{q})$ *using a LR test. If the null hypothesis is not rejected*

3) *Test a VAR*$(\bar{q}-2)$ *against VAR*$(\bar{q}-1)$ *using an LR test. Continue until rejection.*

Clearly, $\bar{q}$ depends on the frequency of the data. For annual data $\bar{q} = 3$; for quarterly data $\bar{q} = 8$; and for monthly data $\bar{q} = 18$ are typical choices. Note that with a sequential approach each null hypothesis is tested conditional on all the previous ones being true and that the chosen q crucially depends on the significance level. Furthermore, when a sequential procedure is used it is important to distinguish between the significance level of individual tests and the significance level of the procedure as a whole - in fact, rejection of a VAR$(\bar{q}-j)$ implies that all VAR$(\bar{q}-j')$ will also be rejected, $\forall j' > j$.

Example 4.7 *Choose as a significance level* 0.05 *and set* $\bar{q} = 6$. *Then a likelihood ratio test for q=5 vs. q=6 has significance level* $1-0.95 = 0.05$. *Conditional on choosing q=5, a test for q=4 vs. q=5 has a significance level* $1-(0.95)^2 = 0.17$ *and the significance level at the* $j-th$ *stage is* $1-(1-.05)^j$. *Hence, if we expect the model to have three or four lags, we better adjust the significance level so that at the second or third stage of the testing, the significance is around* 0.05.

Exercise 4.14 *A LR test restricts each equation to have the same number of lags. Is it possible to choose different lag lengths in different equations? How would you do this in a bivariate VAR?*

While popular, LR tests are unsatisfactory lag selection approaches when the VAR is used for forecasting. This is because LR tests look at the in-sample fit of models (see equation 4.10). When forecasting one would like to have lag selection methods which minimize the (out-of-sample) forecast error. Let $y_{t+\tau} - y_t(\tau)$ be the τ-step ahead forecast error based on time t information and let $\Sigma_y(\tau) = E[y_{t+\tau} - y_t(\tau)][y_{t+\tau} - y_t(\tau)]'$ be its mean square error (MSE). When $\tau = 1$, $\Sigma_y(1) \approx \frac{T+mq}{T}\Sigma_e$ where Σ_e is the variance covariance matrix of the innovations (see e.g. Lutkepohl (1991, p.88)). The next three information criteria choose lag length using transformations of $\Sigma_y(1)$.

- Akaike Information criterion (AIC) : $\min_q AIC(q) = \ln|\Sigma_y(1)|(q) + \frac{2qm^2}{T}$.

- Hannan and Quinn criterion (HQC): $\min_q HQC(q) = \ln|\Sigma_y(1)|(q) + \frac{2qm^2}{T}\ln(\ln T)$.
- Schwarz criterion (SWC): $\min_q SC(q) = \ln|\Sigma_y(1)|(q) + \frac{2qm^2}{T}\ln T$.

All criteria add a penalty to the one-step ahead MSE which depends on the sample size T, the number of variables m and the number of lags q. While for large T penalty differences are unimportant, this is not the case when T is small, as shown in table 4.1.

Criterion	T=40, m=4			T=80, m=4			T=120, m=4			T=120, m=4		
	q=2	q=4	q=6	q=2	q=4	q=6	q=2	q=4	q=6	q=2	q=4	q=6
AIC	0.4	3.2	4.8	0.8	1.6	2.4	0.53	1.06	1.6	0.32	0.64	0.96
HQC	0.52	4.17	6.26	1.18	2.36	3.54	0.83	1.67	2.50	0.53	1.06	1.6
SWC	2.95	5.9	8.85	1.75	3.5	5.25	1.27	2.55	3.83	0.84	1.69	2.52

Table 4.1: Penalties of Akaike, Hannan and Quinn and Schwarz criteria

In general, for $T \geq 20$ SWC and HQC will always choose smaller models than AIC.

The three criteria have different asymptotic properties. AIC is inconsistent (in fact, it overestimates the true order with positive probability) while HQC and SWC are consistent and when $m > 1$, they are both strongly consistent (i.e. they will choose the correct model almost surely). Intuitively, AIC is inconsistent because the penalty function used does not simultaneously goes to infinity as $T \to \infty$ and to zero when scaled by T. Consistency however, it is not the only yardstick to use since consistent methods may have poor small sample properties. Ivanov and Kilian (2001) extensively study the small sample properties of these three criteria using a variety of data generating processes and data frequencies and found that HQC is best for quarterly and monthly data, both when y_t is covariance stationary and when it is a near-unit root process.

Example 4.8 *Consider a quarterly VAR model for the Euro area for the sample 1980:1-1999:4 (T=80); restrict $m = 4$ and use output, prices, interest rates and money (M3) as variables. A constant is eliminated previous to the search. We set $\bar{q} = 7$. Table 4.2 reports the sequential p-values of basic and modified LR tests (first two columns) and the values of the AIC, HQC, SWC criteria (other three columns).*

Different tests select somewhat different lag length. The LR tests select 7 lags but the p-values are non-monotonic and it matters what $\bar{q}$ is. For example, if $\bar{q} = 6$, LR^c selects two lags. Nonmonotonicity appears also for the other three criteria. In general, SWC, which uses the harshest penalty, has a minimum at 1; HQC and AIC have a minimum at 2. Based on these outcomes, we tentatively select a VAR(2).

4.2.2 Lag Length 2

The Wold theorem implies, among other things, that VAR residuals must be white noise. A LR test can therefore be interpreted as a diagnostic to check whether residuals satisfy

Hypothesis	*LR*	*LR^c*	*AIC*	*HQC*	*SWC*
q=6 vs. q=7	*2.9314e-05*	*0.0447*	*-7.5560*	*-6.3350*	*-4.4828*
q=5 vs. q=6	*3.6400e-04*	*0.1171*	*-7.4139*	*-6.3942*	*-4.8514*
q=4 vs. q=5	*0.0509*	*0.5833*	*-7.4940*	*-6.6758*	*-5.4378*
q=3 vs. q=4	*0.0182*	*0.4374*	*-7.5225*	*-6.9056*	*-5.9726*
q=2 vs. q=3	*0.0919*	*0.6770*	*-7.6350*	*-7.2196*	*-6.5914*
q=1 vs. q=2	*3.0242e-07*	*6.8182e-03*	*-7.2266*	*-7.0126*	*-6.6893*

Table 4.2: Lag length of a VAR

this property. Similarly, AIC, HQC and SWC can be seen as trading-off the white noise assumption on the residuals with the best possible out-of-sample forecasting performance.

Another class of tests to lag selection directly examines the properties of VAR residuals. Let $ACRF_e(\tau)^{i,i'}$ denote the cross correlation of e_{it} and $e_{i't}$ at lag $\tau = \ldots, -1, 0, 1, \ldots$. Then , under the null of white noise $ACRF_e(\tau)^{i,i'} = \frac{ACF_e(\tau)^{i,i'}}{\sqrt{ACF_e(0)^{i,i} ACF_e(0)^{i',i'}}} \rightarrow \mathsf{N}(0, \frac{1}{T})$ for each τ (see e.g. Lutkepohl (1991, p.141).

Exercise 4.15 *Design a test for the joint hypothesis that $ACRF_e(\tau) = 0 \ \forall i, i', \tau$ fixed.*

Care must be exercised in implementing white noise tests sequentially - say, starting from an upper $\bar{q}$, checking if the residual are white noise and, if they are, decrease $\bar{q}$ by one value at the time until the null hypothesis is rejected. Since serial correlation is present in incorrectly specified VARs, one must choose a $\bar{q}$ for which the null hypothesis is satisfied.

Exercise 4.16 *Provide a test statistic for the null that $ACRF_e(\tau)^{i,i'} = 0, \ \forall \tau$ which is robust to the presence of heteroschedasticity in VAR residuals.*

In implementing white noise tests, one should remember that since VAR residuals are estimated, the asymptotic covariance matrix of the ACRF must include parameter uncertainty. Contrary to what one would expect, the covariance matrix of the estimated residuals is smaller than the one based on the true ones (see e.g. Lutkepohl (1991, p.142-148)). Hence, $\frac{1}{T}$ is conservative in the sense that the null hypothesis will be rejected less often than indicated by the significance level.

Portmanteau or Q-tests for the whiteness of the residuals can also be used to choose the lag length of a VAR. Both Portmanteau and Q-tests are designed to verify the null that $ACRF_e^\tau = (ACRF_e(1), \ldots, ACRF_e(\tau)) = 0$, (the alternative is $ACRF_e^\tau \neq 0$). The Portmanteau statistic is $PS(\tau) = T \sum_{i=1}^{\tau} tr(ACF(i)' ACF(0)^{-1'} ACF(i) ACF(0)^{-1}) \xrightarrow{D} \chi^2(m^2(\tau - q))$ for $\tau > q$ under the null. The Q-statistic is $QS(\tau) = T(T+2) \sum_{i=1}^{\tau} \frac{1}{T-i}$ $tr(ACF(i)' ACF(0)^{-1'} ACF(i) ACF(0)^{-1})$. For large T, it has the same asymptotic distribution as $PS(\tau)$.

Exercise 4.17 *Use US quarterly data from 1960:1 to 2002:4 to optimally select the lag length of a VAR with output, prices, nominal interest rate and money. Use modified LR,*

AIC, HQC, SWC and white noise tests. Does it make a difference if the sample is 1970-2003 or 1980-2003? How do you interpret differences across tests and/or samples?

4.2.3 Nonlinearities and nonnormalities

So far we have focused on linear specifications. Since time aggregation washes most of the nonlinearities out, the focus is hardly restrictive, at least for quarterly data. However, with monthly data nonlinearities could be important (especially if financial data is used). Furthermore, time variations in the coefficients (see chapter 10), outliers or structural breaks may also generate (in a reduced form sense) nonlinearities and nonnormalities in the residuals of a constant coefficient VAR. Hence, one wants methods to detect departures from nonlinearities and nonnormalities if they exist.

In deriving the MA representation we have used linear projections. Since omitted nonlinear terms will end up in the error term, the same ideas employed in testing for white noise residuals can be used to check if nonlinear effects are present.

Two ways of formally testing for nonlinearities are the following: i) run a regression of estimated VAR residuals on nonlinear functions of the lagged dependent variables and examine the significance of estimated coefficients adjusting standard errors for the fact that e_t is proxied by estimated residuals. ii) Directly insert high order terms in the VAR and examine their significance. Graphical techniques, e.g. a scatter plot of estimated residuals against nonlinear functions of the regressors, could also be used as diagnostics.

There is also an indirect approach to check for nonlinearities which builds on the idea that whenever nonlinear terms are important, the moments of the residuals have a special structure. In particular, their distribution will be non-normal, even in large samples.

Testing for nonnormalities is simple: a normal white noise process with unit variance has zero skewness (third moment) and kurthosis (the forth moment) equal to 3. Hence, an asymptotic test for nonnormalities is as follows. Let $\hat{e}_t = y_t - \sum_j \hat{A}_j y_{t-j}$; $\Sigma_e = \frac{1}{T-1}\sum_t \hat{e}_t \hat{e}_t'$; $\tilde{e}_t = \tilde{\mathcal{P}}^{-1}\hat{e}_t$; $\tilde{\mathcal{P}}\tilde{\mathcal{P}}' = \Sigma_e$ where $\hat{A}_j$ is an estimator of A_j. Define $\mathsf{S}_{1i} = \frac{1}{T}\sum_t \tilde{e}_{it}^3$; $\mathsf{S}_{2i} = \frac{1}{T}\sum_t \tilde{e}_{it}^4, i = 1, \ldots, m$, $\mathsf{S}_j = (\mathsf{S}_{j1}, \ldots, \mathsf{S}_{jm})', j = 1, 2$ and let 3_m be a $m \times 1$ vector with 3 in each entry. Then $\sqrt{T}\begin{bmatrix} \mathsf{S}_1 \\ \mathsf{S}_2 - 3_m \end{bmatrix} \xrightarrow{D} \mathsf{N}(0, \begin{bmatrix} 6 \times I_m & 0 \\ 0 & 24 \times I_m \end{bmatrix})$.

4.2.4 Stationarity

Covariance stationarity is crucial to derive a VAR representation with constant coefficients. However, a time varying MA representation for a nonstationary y_t always exists if the other assumptions used in the Wold theorem hold. If $\sum_j D_{jt}^2 < \infty$ for all t, a non-stationary VAR representation can be derived. Hence, time varying coefficient VAR models, which we examine in chapter 10, are the natural alternative to covariance stationary structures.

While covariance stationarity is unnecessary, it is a convenient property to have when estimating VAR models. Also, although models with smooth changes in the coefficients may be the natural extensions of covariance stationary models, the literature has focused on a more extreme form of nonstationarity: unit root processes. Unit root models are

less natural for two reasons: they imply drastically different dynamic properties; classical statistics has difficulties in testing this null hypothesis in the presence of a near-unit root alternative (see e.g. Watson (1995)). Despite these problems, contrasting stationary vs. unit root behavior has become a rule, the common wisdom being that macroeconomic time series are characterized by near-unit root behavior, i.e. they are in the grey area where the tests have low power. Hence, it will take a long time for a randomly perturbed series to revert back to the original (steady) state.

Unit root tests are somewhat tangential to the scope of the book. Favero (2001) provides an excellent review of this literature. Hence, we limit attention to the implications that nonstationary (or near nonstationary) has for the specification of the VAR, for the estimation of the parameters and for the identification of structural shocks.

If a test has detected one or more unit roots, how should one proceed in specifying a VAR? Suppose we are confident in the testing results and that all variables are either stationary or integrated, but no cointegration is detected. Then one would difference unit-root variables until covariance stationary is obtained and estimate the VAR using transformed variables. For example, if all variables are I(1), a VAR in growth rates is appropriate.

Specification is simple also when there are some cointegrating relationships. For example, both prices and money may display unit root behavior but real balances may be stationary. In this case, one typically transforms the VAR into a vector error correction model (VECM) and either imposes the cointegrating relationships (using the theoretical or the estimated restrictions) or jointly estimates short run and long run coefficients from the data. VECMs are preferable here to differenced VARs because the latters throw away information about the long run properties of the data. Plugging-in estimates of the long run relationships is justified since estimates of the long run relationships are super-consistent, i.e. they asymptotically converge at the rate T (estimates of short run relationships converge asymptotically at the rate $T^{0.5}$). Since a VECM is a reparametrization of the VAR in levels, the latter is appropriate if all variables are cointegrated, even though some (or all) of its components are not covariance stationary.

Despite two decades of work in the area, unit root tests still have poor small sample properties. Furthermore, barring exceptional circumstances, neither explosive nor unit root behavior has been observed in long stretches of OECD macroeconomic data. Both reasons may cast doubts about the non-stationarities detected and the usefulness of such tests.

When doubts about the tests exist, one can indirectly check the reasonableness of the stationarity assumption by studying estimated residuals. In fact, if y_t is nonstationary and no cointegration emerges, the estimated residuals are likely to display nonstationary path. Hence a plot of the VAR residuals may indicate a problem if it exists. Practical experience suggests that VAR residuals show breaks and outliers but they rarely display unit root type behavior. Hence, a level VAR could be appropriate even when y_t looks nonstationary. It is also important to remember that the properties of y_t are important in testing hypotheses about the coefficients since classical distribution theory is different when unit roots are present. Consistent estimates of VAR coefficients obtain with classical methods even when unit roots are present (see Sims, Stock and Watson (1990)).

A final argument against the use of specification tests for stationarity comes from a Bayesian perspective. In Bayesian analysis the posterior distribution of the quantities of interest is all that matters. While Bayesian and classical analyses have many common aspects, they dramatically differ when unit roots are present. In particular, while the classical asymptotic distribution of coefficients estimates under unit roots is nonstandard, the posterior distribution is unchanged. Therefore, if one takes a Bayesian perspective to testing, no adjustment for nonstationarity is required.

Finally, one should remember that pretesting has consequences for the distribution of parameters estimates since incorrect choices produce inconsistent estimates of the quantities of interest. To minimize pretesting problems, we recommend to start assuming covariance stationarity and deviate from it only if the data overwhelmingly suggests the opposite.

4.2.5 Breaks

While exact unit root behavior is unlikely to be relevant in macroeconomics, changes in the intercept, in the dynamics or in the covariance matrix of a vector of time series are quite common. A time series with breaks is neither stationary nor covariance stationary. To avoid problems, applied researchers typically focus attention on subsamples which are (assumed to be) homogenous. However, this is not always possible: the break may occur at the end of the sample (e.g. creation of the Euro); there maybe several of them; or they may be linked to expansions and contractions and it may be unwise to throw away runs with these characteristics.

While structural breaks with dramatic changing dynamics may sometimes occur (e.g. breakdown or unification of a country), it is more often the case that time series display slowly evolving features with no abrupt changes at one specific point - a pattern which would be more consistent with a time varying coefficient specifications. Nevertheless, it may be useful to have tools to test for structural breaks if visual inspection suggests that such a pattern may be present. If the break date is known, Chow tests can be used. Let Σ_e^{re} be the covariance matrix of the VAR residuals with no breaks and $\Sigma_e^{un} = \Sigma_e^{un}(1,\bar{t}) + \Sigma_e^{un}(\bar{t}+1,T)$ is the covariance matrix when a break is allowed at $\bar{t}$. Then $CS(\bar{t}) = \frac{|\Sigma_e^{re}|-|\Sigma_e^{un}|)/\nu}{|\Sigma_e^{un}|/T-\nu} \sim F(\nu, T-\nu)$ where ν is the number of regressors in the model. When $\bar{t}$ is unknown but suspected to occur within an interval, one could run Chow tests for all $\bar{t} \in [t_1, t_2]$, take $\max_{\bar{t}} CT(\bar{t})$ and compare it with a modified F-distribution (critical values are e.g. in Stock and Watson (2002, p. 111)).

An alternative testing approach can be obtained by noting that if no break occurs the τ-steps ahead forecast error of $y_{t+\tau}$, $e_t(\tau) = y_{t+\tau} - y_t(\tau)$, should be similar to sample residuals. Then, under the null of no breaks at forecasting horizon τ, τ large $e_t(\tau) \xrightarrow{D} \mathsf{N}(0, \Sigma_e(\tau))$.

Exercise 4.18 *Show that an appropriate statistic to check for breaks over τ forecasting horizons is $FT(\tau) = e_t \Sigma_e^{-1} e_t \xrightarrow{D} \chi^2(\tau)$ under the null of no breaks, T large, where $e_t = (e_t(1), \dots e_t(\tau))$. (The alternative here is that the DGP for y_t differs before and after t).*

As usual these tests may be biased in small samples. A small sample version of the forecasting test is obtained using $\Sigma_e^c(\tau) = \Sigma_e(\tau) + \frac{1}{T}E[\frac{\partial y_t(\tau)}{\partial \alpha'}\Sigma_\alpha \frac{\partial y_t(\tau)}{\partial \alpha}']$ in place of $\Sigma_e(\tau)$.

4.3 Alternative Representations of VAR(q)

There are two alternative representations for a $VAR(q)$ which are easier to manipulate than (4.7) and are of use when deriving estimators of the unknown parameters of the model.

4.3.1 Companion form representation

The companion form representation transforms a VAR(q) model in a larger scale VAR(1) model and it is useful when one needs to compute moments or derive parameter estimates.

Let $\mathsf{Y}_t = \begin{bmatrix} y_t \\ y_{t-1} \\ \dots \\ y_{t-q+1} \end{bmatrix}$; $\mathsf{E}_t = [\begin{bmatrix} e_t \\ 0 \\ \dots \end{bmatrix} 0$; $\mathsf{A} = \begin{bmatrix} A_1 & A_2 & \dots & A_q \\ I_m & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & \dots & I_m & 0 \end{bmatrix}$. Then (4.7) is

$$\mathsf{Y}_t = \mathsf{A}\mathsf{Y}_{t-1} + \mathsf{E}_t \quad \mathsf{E}_t \sim (0, \Sigma_E) \tag{4.11}$$

where $\mathsf{Y}_t, \mathsf{E}_t$ are $mq \times 1$ vectors and A is $mq \times mq$ matrix.

Example 4.9 *Consider a bivariate VAR(2) model. Here* $\mathsf{Y}_t = [y_t, y_{t-1}]'$ $\mathsf{E}_t = [e_t, 0]'$*, are a* 4×1 *vectors, and* $\mathsf{A} = \begin{bmatrix} A_1 & A_2 \\ I_2 & 0 \end{bmatrix}$ *is a* 4×4 *matrix.*

Moments of y_t can be immediately calculated from (4.11).

Example 4.10 *The unconditional mean of* y_t *can be computed using* $E(\mathsf{Y}_t) = [(I - \mathsf{A}\ell)^{-1}]$ $E(\mathsf{E}_t) = 0$ *and a selection matrix which picks the first* m *elements out of* $E(\mathsf{Y}_t)$*. To calculate the unconditional variance notice that, because of covariance stationarity*

$$\begin{aligned} E[(\mathsf{Y}_t - E(\mathsf{Y}_t))(\mathsf{Y}_t - E(\mathsf{Y}_t))'] &= \mathsf{A}E_t[(\mathsf{Y}_{t-1} - E(\mathsf{Y}_{t-1}))(\mathsf{Y}_{t-1} - E(\mathsf{Y}_{t-1})'] \mathsf{A}' + \Sigma_E \\ \Sigma_Y &= \mathsf{A}\Sigma_Y \mathsf{A}' + \Sigma_E \end{aligned} \tag{4.12}$$

To solve (4.12) for Σ_Y *we will make use of the following result.*

Result 4.1 *If* T, V, R *are conformable matrices,* $vec(TVR) = (R' \otimes T)vec(V)$*.*

Then $vec(\Sigma_Y) = [I_{mq} - (\mathsf{A} \otimes \mathsf{A})]^{-1} vec(\Sigma_E)$ *where* I_{mq} *is a* $mq \times mq$ *identity matrix.*

Unconditional covariances and correlations can also be easily computed. In fact

$$\begin{aligned} ACF_Y(\tau) &= E[(\mathsf{Y}_t - E(\mathsf{Y}_t))(\mathsf{Y}_{t-\tau} - E(\mathsf{Y}_{t-\tau})'] \\ &= \mathsf{A}E_t[(\mathsf{Y}_{t-1} - E(\mathsf{Y}_{t-1}))(\mathsf{Y}_{t-\tau} - E(\mathsf{Y}_{t-\tau}))'] + E[\mathsf{E}_t(\mathsf{Y}_{t-\tau} - E(\mathsf{Y}_{t-\tau}))'] \\ &= \mathsf{A}ACF_Y(\tau - 1) = \mathsf{A}^\tau \Sigma_Y \quad \tau = 1, 2, \dots \end{aligned} \tag{4.13}$$

The companion form could also be used to obtain the spectral density matrix of y_t. Let $ACF_E(\tau) = cov(\mathsf{E}_t, \mathsf{E}_{t-\tau})$. Then the spectral density of E_t is $\mathcal{S}_E(\omega) = \frac{1}{2\pi}\sum_{\tau=-\infty}^{\infty} e^{-i\omega\tau} ACF_E(\tau)$ and $vec[\mathcal{S}_Y(\omega)] = [I(\omega) - \mathsf{A}(\omega)\mathsf{A}(-\omega)']vec[\mathcal{S}_E(\omega)]$ where $I(\omega) = \sum_j e^{-i\omega j} I$, $\mathsf{A}(\omega) = \sum_j e^{-i\omega j}\mathsf{A}^j$ and $\mathsf{A}(-\omega)'$ is the complex conjugate of $\mathsf{A}(\omega)$.

Exercise 4.19 *Suppose a VAR(2) has been fitted to unemployment and inflation data and* $\hat{A}_1 = \begin{bmatrix} 0.95 & 0.23 \\ 0.21 & 0.88 \end{bmatrix}$, $\hat{A}_2 = \begin{bmatrix} -0.05 & 0.13 \\ -0.11 & 0.03 \end{bmatrix}$ *and* $\hat{\Sigma}_e = \begin{bmatrix} 0.05 & 0.01 \\ 0.01 & 0.06 \end{bmatrix}$ *have been obtained. Calculate the spectral density matrix of* y_t. *What is the value of* $\mathcal{S}_Y(\omega = 0)$*?*

A companion form representation has also computational advantages when deriving estimators of the unknown parameters of the model. We first consider estimators obtained when no constraints (lag restrictions, zero restrictions, etc.) are imposed on the VAR; when $y_{-q+1}, \ldots y_0$ are fixed and e_t are normally distributed with covariance matrix Σ_e.

Given the VAR structure, $(y_t|y_{t-1}, \ldots, y_0, y_{-1}, \ldots, y_{-q+1}) \sim \mathsf{N}(\mathsf{A}_1\mathsf{Y}_{t-1}, \Sigma_e)$ where A_1 is a $m \times mq$ matrix containing the first m rows of A. The density of y_t is $f(y_t|y_{t-1}, \ldots, \mathsf{A}_1, \Sigma_e) = (2\pi)^{0.5m}|\Sigma_e|^{-0.5} exp[-0.5(y_t - \mathsf{A}_1\mathsf{Y}_{t-1})'\Sigma_e^{-1}(y_t - \mathsf{A}_1\mathsf{Y}_{t-1})]$. Hence $f(y_t, y_{t-1}, \ldots, |\mathsf{A}_1, \Sigma_e) = \prod_{t=1}^{T} f(y_t|y_{t-1}, \ldots, \mathsf{A}_1, \Sigma_e)$ and the log likelihood is

$$\mathcal{L}(\mathsf{A}_1, \Sigma_e|y_t) = -\frac{T}{2}(m\log(2\pi) - \log|\Sigma_e|) - \frac{1}{2}\sum_t (y_t - \mathsf{A}_1\mathsf{Y}_{t-1})'\Sigma_e^{-1}(y_t - \mathsf{A}_1\mathsf{Y}_{t-1})] \quad (4.14)$$

Taking the first order conditions with respect to $vec(\mathsf{A}_1)$ leads to

$$\mathsf{A}'_{1,ML} = [\sum_{t=1}^{T} \mathsf{Y}_{t-1}\mathsf{Y}'_{t-1}]^{-1}[\sum_{t=1}^{T} \mathsf{Y}_{t-1}y'_t] = \mathsf{A}'_{1,OLS} \quad (4.15)$$

Hence, when no restrictions are imposed, ML and OLS estimators of the first m rows of the companion matrix A coincide. Note that an estimator of the j-th row of A_1 (an $1 \times mq$ vector) is $\mathsf{A}'_{1j} = [\sum_t \mathsf{Y}_{t-1}\mathsf{Y}'_{t-1}]^{-1}[\sum_{t=1}^{T} \mathsf{Y}_{t-1}y_{jt}]$.

Exercise 4.20 *Provide conditions for* $\mathsf{A}_{1,ML}$ *to be consistent. Is it efficient?*

Exercise 4.21 *Show that if there are no restrictions on the VAR, OLS estimation of the parameters, equation by equation, is consistent and efficient.*

The result of exercise 4.21 is important: as long as all variables appear with the same lags in every equation, single equation OLS estimation is sufficient. Intuitively, such a VAR is a seemingly unrelated regression (SUR) model and for such models single equation and system wide methods are equally efficient (see e.g. Hamilton (1994, p.315)).

Using $\mathsf{A}_{1,ML}$ into the log likelihood we obtain $\ln\mathcal{L}(\Sigma_e|y_t) = -\frac{Tm}{2}\ln(2\pi) - \frac{T}{2}\ln|\Sigma_e|) - \frac{1}{2}\sum_{t=1}^{T} e'_{t,ML}\Sigma_e^{-1}e_{t,ML}$ where $e_{t,ML} = (y_t - \mathsf{A}_{1,ML}\mathsf{Y}_{t-1})$. Taking the first order conditions

with respect to $vech(\Sigma_e)$, where $vech(\Sigma_e)$ vectorizes the symmetric matrix Σ_e, and using the fact that $\frac{\partial(b'\mathcal{Q}b)}{\partial\mathcal{Q}} = b'b$; $\frac{\partial \log|\mathcal{Q}|}{\partial\mathcal{Q}} = (\mathcal{Q}')^{-1}$ we have $\frac{T}{2}\Sigma_e' - \frac{1}{2}\sum_{t=1}^{T} e_{t,ML}e_{t,ML}' = 0$ or

$$\Sigma_{ML}' = \frac{1}{T}\sum_{t=1}^{T} e_{t,ML}e_{t,ML}' \tag{4.16}$$

and the ML estimate of the (i,i') element of Σ_e is $\sigma_{i,i'} = \frac{1}{T}\sum_{t=1}^{T} e_{it,ML}e_{i't,ML}'$.

Exercise 4.22 *Show that Σ_{ML} is biased but consistent.*

4.3.2 Simultaneous equations format

Two other useful transformations of a VAR are obtained using the format of a simultaneous equations system. The first is obtained setting $x_t = [y_{t-1}, y_{t-2}, \ldots]$; $\mathsf{X} = [x_1, \ldots, x_T]'$ (a $T \times mq$ matrix), $\mathsf{Y} = [y_1, \ldots, y_T]'$ (a $T \times m$ matrix) and letting $\mathsf{A} = [A_1', \ldots A_q']' = \mathsf{A}_1'$ be a $mq \times m$ matrix to have

$$\mathsf{Y} = \mathsf{X}\mathsf{A}' + \mathsf{E} \tag{4.17}$$

The second transformation is obtained from (4.17). The equation for variable i in fact is $\mathsf{Y}_i = \mathsf{X}\mathsf{A}_i + \mathsf{E}_i$. Stacking the columns of $\mathsf{Y}_i, \mathsf{E}_i$ into $mT \times 1$ vectors we have

$$y = (I_m \otimes \mathsf{X})\alpha + e \equiv X\alpha + e \tag{4.18}$$

Note that in (4.17) all variables are grouped together for each t; in (4.18) all time periods for one variable are grouped together. As shown in chapter 10, (4.18) is useful to decompose the likelihood function of a VAR(q) into the product of a normal density, conditional on the OLS estimates of the VAR parameters, and a Wishart density for Σ_e^{-1}.

Using these representations it is immediate to compute moments of y_t.

Example 4.11 *The unconditional mean of y_t is $E(\mathsf{Y}) = E(\mathsf{X})\mathsf{A}'$ or $E(y) = E(I_m \otimes \mathsf{X})\alpha$. The unconditional variance is $E[\mathsf{Y}] \equiv \Sigma_Y = E\{[\mathsf{X} - E(\mathsf{X})]\mathsf{A}' - \mathsf{E}\}^2$ or $\Sigma_Y = E\{[(I_m \otimes \mathsf{X}) - E(I_m \otimes \mathsf{X})]\alpha + e\}^2$.*

Exercise 4.23 *Using (4.18), assuming that $\Sigma_{xx} = p\lim\frac{X'X}{T}$ exists and is non-singular and $\frac{1}{\sqrt{T}}vec(Xe) \xrightarrow{D} \mathsf{N}(0, \Sigma_{xx} \otimes \Sigma_e)$ show: (i)$p\lim_{T\to\infty}\alpha_{OLS} = \alpha$; (ii) $\sqrt{T}(\alpha_{OLS} - \alpha) \xrightarrow{D} \mathsf{N}(0, \Sigma_{xx}^{-1} \otimes \Sigma_e)$; (iii) $\Sigma_{e,OLS} = \frac{(y-X\alpha)(y-X\alpha)'}{T-mq}$ is such that $p\lim\sqrt{T}(\Sigma_{e,OLS} - \frac{ee'}{T}) = 0$.*

Estimators of the VAR parameters can also be obtained via the Yule-Walker equations. From (4.7) we have that $E[(y_t - E(y_t))(y_{t-\tau} - E(y_{t-\tau}))] = A(\ell)E[(y_{t-1} - E(y_{t-1}))(y_{t-\tau} - E(y_{t-\tau}))] + E[e_t(y_{t-\tau} - E(y_{t-\tau}))]$ for all $\tau \geq 0$. Hence, letting $ACF_y(\tau) = E[(y_t - E(y_t))(y_{t-\tau} - E(y_{t-\tau}))]$ we have

$$ACF_y(\tau) = A_1 ACF_y(\tau-1) + A_2 ACF_y(\tau-2) + \ldots + A_q ACF_y(\tau-q) \tag{4.19}$$

Example 4.12 *If $q = 1$ (4.19) reduces to $ACF_y(\tau) = A_1 ACF_y(\tau - 1)$. Given estimates of A_1 and Σ_e, we have that $ACF_y(0) \equiv \Sigma_y = A_1 \Sigma_y A_1' + \Sigma_e$ so $vec(\Sigma_y) = (I - A_1 \otimes A_1) vec(\Sigma_e)$ and $ACF_y(1) = A_1 ACF_y(0)$, $ACF_y(2) = A_1 ACF_y(1)$, etc.*

Equation (4.19) can also be more compactly written as $ACF_y = \mathsf{A}_1 ACF_y^*$ where $ACF_y = [ACF_y(1), \ldots ACF_y(q)]$; and $ACF_y^* = \begin{bmatrix} ACF_y(0) & \ldots & ACF_y(q-1) \\ \ldots & \ldots & \ldots \\ ACF_y(-q+1) & \ldots & ACF_y(0) \end{bmatrix}$. Then an estimate of A_1 is $\mathsf{A}_{1,YW} = ACF_y (ACF_y^*)^{-1}$.

Exercise 4.24 *Show that $\mathsf{A}_{1,YW} = \mathsf{A}_{1,ML}$. Conclude that Yule-Walker and ML estimators have the same asymptotic properties.*

Exercise 4.25 *Show how to modify the Yule-Walker estimator when $E(y_t)$ is unknown. Show that the resulting estimator is asymptotically equivalent to $\mathsf{A}_{1,YW}$.*

It is interesting to study what happens when a VAR is estimated under some restrictions (exogeneity, cointegration, lag elimination, etc.). Suppose restrictions are of the form $\alpha = R\theta + r$ where R is $mk \times k_1$ matrix of rank k_1; r is a $mk \times 1$ vector; θ a $k_1 \times 1$ vector.

Example 4.13 *i) Consider the restriction $A_q = 0$. Here $k_1 = m^2(q-1)$, $r = 0$, and $R = [I_{k_1}, 0]$*
ii) Suppose that y_{2t} is exogenous for y_{1t} in a bivariate VAR(2). Here $R = blockdiag[R_1, R_2]$ where R_i, $i = 1, 2$ is upper triangular.

Using (4.18) we have $y = (I_m \otimes \mathsf{X})\alpha + e = (I_m \otimes \mathsf{X})(R\theta + r) + e$ or $y - (I_m \otimes \mathsf{X})r = (I_m \otimes \mathsf{X})R\theta + e$. Since $\frac{\partial \ln \mathcal{L}}{\partial \theta} = R \frac{\partial \ln \mathcal{L}}{\partial \alpha}$ then

$$\theta_{ML} = [R'(\Sigma_e^{-1} \otimes \mathsf{X}'\mathsf{X})R]^{-1} R'[\Sigma_e^{-1} \otimes \mathsf{X}](y - (I_m \otimes \mathsf{X})r) \quad (4.20)$$

$$\alpha_{ML} = R\,\theta_{ML} + r \quad (4.21)$$

$$\Sigma_e = \frac{1}{T} \sum_t e_{ML} e_{ML}' \quad (4.22)$$

Exercise 4.26 *Verify that when a VAR is estimated under some restrictions:*
i) ML estimates are different from OLS estimates.
ii) ML estimates are consistent and efficient if the restrictions are true but inconsistent if the restrictions are false.
iii) OLS is consistent when stationarity is incorrectly assumed but t-tests are incorrect.
iv) OLS is inconsistent if lag restrictions are incorrect.

4.4 Reporting VAR results

It is rare to report estimated VAR coefficients. Since the number of parameters is large presenting all of them is cumbersome. Furthermore, they are poorly estimated: except

for the first own lag, in general, they are all insignificant. It is therefore typical to report functions of the VAR coefficients which summarize information better, have some economic meaning and, hopefully, are more precisely estimated. Among the many possible functions, three are typically used: impulse responses, variance and historical decompositions. Impulse responses trace out the MA of the system, i.e. they describe how $y_{it+\tau}$ responds to a shock in $e_{i't}$; the variance decomposition measures the contribution of $e_{i't}$ to the variability of $y_{it+\tau}$; the historical decomposition describes the contribution of shock $e_{i't}$ to the deviations of $y_{it+\tau}$ from its baseline forecasted path.

4.4.1 Impulse responses

There are three ways to calculate impulse responses which roughly correspond to recursive, nonrecursive (companion form) and forecast revision approaches. In the recursive approach, the impulse response matrix at horizon τ is $D_\tau = \sum_{j=1}^{max[\tau,q]} D_{\tau-j}A_j$ where $D_0 = I,\ A_j = 0\ \forall\ \tau \geq q$. Clearly, a consistent estimate is obtained if a consistent $\hat{A}_j$ is used in place of A_j.

Example 4.14 *Consider a VAR(2) with $y_t = A_0 + A_1 y_{t-1} + A_2 y_{t-2} + e_t$. Then the response matrices are: $D_0 = I,\ \ D_1 = D_0 A_1,\ \ D_2 = D_1 A_1 + D_0 A_2,\ \ldots, D_\tau = D_{\tau-1}A_1 + D_{\tau-2}A_2$.*

Calculation of meaningful impulse responses requires orthogonal disturbances. Let $\tilde{\mathcal{P}}$ be a square matrix such that $\tilde{\mathcal{P}}\tilde{\mathcal{P}}' = \Sigma_e$. Then the impulse response matrix to orthogonal shocks $\tilde{e}_t = \tilde{\mathcal{P}}^{-1}e_t$ at horizon τ is $\tilde{D}_\tau = D_\tau\tilde{\mathcal{P}}$.

Exercise 4.27 *Provide the first 5 elements of the MA representation of a bivariate VAR(3) with orthogonal shocks.*

When the VAR is in a companion form, we can compute impulse responses in a different way. Using (4.11) and repeatedly substituting for $\mathsf{Y}_{t-\tau},\ \tau = 1, 2, \ldots$ we have:

$$\mathsf{Y}_t = \mathsf{A}^t\mathsf{Y}_0 + \sum_{\tau=0}^{t-1}\mathsf{A}^\tau\mathsf{E}_{t-\tau} \tag{4.23}$$

$$= \mathsf{A}^t\mathsf{Y}_0 + \sum_{\tau=0}^{t-1}\tilde{\mathsf{A}}^\tau\tilde{\mathsf{E}}_{t-\tau} \tag{4.24}$$

where $\tilde{\mathsf{A}}^\tau = \mathsf{A}^\tau\tilde{\mathsf{P}},\ \tilde{\mathsf{E}}_{t-\tau} = \tilde{\mathsf{P}}^{-1}\mathsf{E}_t,\ \tilde{\mathsf{P}}\tilde{\mathsf{P}}' = \Sigma_E$. (4.23) is used with non-orthogonal residuals, (4.24) with orthogonal ones. The first m rows of A^τ provide the required responses.

Exercise 4.28 *Using the companion form of a bivariate VAR(2) show the first 4 elements of A^τ.*

A final way to compute impulse responses uses forecast revisions of future y_ts. We will use the companion form representation to illustrate the point but the argument goes

through with any representation. Let $\mathsf{Y}_t(\tau) = \mathsf{A}^\tau \mathsf{Y}_t$ and $\mathsf{Y}_{t-1}(\tau) = \mathsf{A}^{\tau+1}\mathsf{Y}_{t-1}$ be the τ-steps and $\tau+1$-steps ahead forecast of Y_t. Hence the forecast revision is

$$Rev_t(\tau) = \mathsf{Y}_t(\tau) - \mathsf{Y}_{t-1}(\tau) \quad = \quad \mathsf{A}^\tau[\mathsf{Y}_t - \mathsf{A}\mathsf{Y}_{t-1}] = \mathsf{A}^\tau \mathsf{E}_t \tag{4.25}$$

Example 4.15 *Suppose we shock the i'-th component of e_t once at time t, i.e. $e_{i't} = 1$; $e_{i'\tau} = 0$, $\tau > t$; $e_{it} = 0\ \forall i \neq i', \forall t$. Then $Rev_{t,i'}(1) = \mathsf{A}_{i',.}$; $Rev_{t,i'}(2) = \mathsf{A}^2_{i',.}$; $Rev_{t,i'}(\tau) = \mathsf{A}^\tau_{i',.}$ where $\mathsf{A}_{i',.}$ is the i-th column of A. Therefore, the response of $y_{i,t+\tau}$ to a shock in $e_{i't}$ can be read off the τ-step ahead forecast revisions.*

Example 4.16 *At times cumulative multipliers are required. For example, in examining the effects of fiscal disturbances on output one may want to measure the cumulative displacement produced by a shock up to horizon τ. Alternatively, in examining the relationship between money growth and inflation one may want to know whether an increase in the former translates in an increase in the latter in the long run of the same amount. In the first case one computes $\sum_{j=0}^{\tau} D_j$, in the second $\lim_{j\to\infty} \sum_{j=0}^{\tau} D_j$.*

4.4.2 Variance decomposition

To derive the variance decomposition we use (4.7). The τ-step ahead forecast error is $y_{t+\tau} - y_t(\tau) = \sum_{j=0}^{\tau-1} \tilde{D}_j \tilde{e}_{t+\tau-j}$ where $D_0 = I$ and $\tilde{e}_t = \tilde{\mathcal{P}}^{-1} e_t = \tilde{\mathcal{P}}_1^{-1} e_{1t} + \ldots + \tilde{\mathcal{P}}_m^{-1} e_{mt}$ are orthogonal disturbances. Hence $\Sigma_{\tilde{e}} = \tilde{\mathcal{P}}_1^{-1}\tilde{\mathcal{P}}_1^{-1\prime}\Sigma_e + \ldots + \tilde{\mathcal{P}}_m^{-1}\tilde{\mathcal{P}}_m^{-1\prime}\Sigma_e$. The MSE of the forecast is

$$\begin{aligned} MSE(\tau) &= E[y_{t+\tau} - y_t(\tau)]^2 = \Sigma_e + D_1\Sigma_e D_1' + \ldots + D_{\tau-1}\Sigma_e D_{\tau-1}' \\ &= \sum_{i=1}^{m} \Sigma_{\tilde{e}} (\tilde{\mathcal{P}}_i^{-1}\tilde{\mathcal{P}}_i^{-1\prime} + \tilde{D}_1\tilde{\mathcal{P}}_i^{-1}\tilde{\mathcal{P}}_i^{-1\prime}\tilde{D}_1' + \ldots + \tilde{D}_{\tau-1}\tilde{\mathcal{P}}_i^{-1}\tilde{\mathcal{P}}_i^{-1\prime}\tilde{D}_{\tau-1}') \end{aligned} \tag{4.26}$$

Hence the percentage of the variance in $y_{i,t+\tau}$ due to $e_{i',t}$

$$VD_{i,i'}(\tau) = \frac{\Sigma_{\tilde{e}} (\tilde{\mathcal{P}}_{i'}^{-1}\tilde{\mathcal{P}}_{i'}^{-1\prime} + \tilde{D}_{1i}\tilde{\mathcal{P}}_{i'}^{-1}\tilde{\mathcal{P}}_{i'}^{-1\prime}\tilde{D}_{1i}' + \ldots + \tilde{D}_{\tau-1,i}\tilde{\mathcal{P}}_{i'}^{-1}\tilde{\mathcal{P}}_{i'}^{-1\prime}\tilde{D}_{\tau-1,i}')}{MSE(\tau)} \tag{4.27}$$

A compact way to rewrite (4.27)is $VD(\tau) = \Sigma_{D_\tau}^{-1} \sum_{j=0}^{\tau-1} D_j \bigodot D_j$ where $\Sigma_{D_\tau} = diag[\Sigma_{D_\tau,11}, \ldots, \Sigma_{D_\tau,mm}] = \sum_{j=0}^{\tau-1} D_j D_j'$ and where $D_j \bigodot D_j$ is a matrix with $D_j^{i,i'} * D_j^{i,i'}$ in the i,i' position ($\bigodot$ is called Hadamman product (see e.g. Mittnick and Zadrozky(1993)).

4.4.3 Historical decomposition

Let $e_{i,t}(\tau) = y_{i,t+\tau} - y_{i,t}(\tau)$ be the τ-steps ahead forecast error in the i-th variable of the VAR. The historical decomposition of $e_{i,t}(\tau)$ can be calculated using

$$e_{i,t}(\tau) = \sum_{i'=1}^{m} \tilde{D}^{i'}(\ell)\tilde{e}_{i't+\tau} \tag{4.28}$$

Example 4.17 *Consider a bivariate VAR(1). At horizon τ we have $y_{t+\tau} = Ay_{t+\tau-1} + e_{t+\tau} = \ldots = A^\tau y_t + \sum_{j=0}^{\tau-1} A^j e_{t+\tau-j}$ so that $e_t(\tau) = \sum_{j=0}^{\tau-1} A^j e_{t+\tau-j} = A(\ell)e_{t+\tau}$. Hence, deviations from the baseline forecasts of the first variable from t to t+τ due to, say, structural supply shocks are $\tilde{A}_{11}(\ell)\tilde{e}_{1,t+\tau}$ and to, say, structural demand shocks are $\tilde{A}_{12}(\ell)\tilde{e}_{2,t+\tau}$.*

From (4.27) and (4.28) it is immediate to notice that the ingredients needed to compute impulse responses, variance and historical decompositions are the same. Therefore, these statistics package the same information in a different way.

Exercise 4.29 *Using the estimate obtained in exercise 4.19, compute the variance and the historical decomposition for the two variables at horizons 1,2 and 3.*

4.4.4 Distribution of Impulse Responses

To assess the statistical (and the economic) significance of the effect of certain shocks, we need standard errors. As we have shown, impulse responses, variance and historical decompositions are complicated functions of the estimated VAR coefficients and of the covariance matrix of the shocks. Therefore, even when the distribution of the latters is known, it is not easy to find their distribution. In this subsection we describe three approaches to compute standard errors: one based on asymptotic theory and two based on resampling methods. All procedures are easy to implement when orthogonal shocks are generated by Choleski factorizations, i.e. if $\tilde{P}$ is lower triangular and need minor modification when the system is not contemporaneously recursive (but just-identified). In the other cases, resampling methods have a slight computational hedge.

Since impulse responses, variance and historical decompositions all use the same information we only discuss how to compute standard errors for impulse responses. The reader will be asked to derive the corresponding expressions for the other two statistics.

•The δ-method

The method pioneered by (Lutkepohl (1991)) and Mittnick and Zadrozky (1993) uses asymptotic approximations and works as follows. Suppose that $\alpha \xrightarrow{D} \mathsf{N}(0, \Sigma_\alpha)$. Then any differentiable function $f(\alpha)$ will have asymptotically the distribution $\mathsf{N}(0, \frac{\partial f}{\partial \alpha}\Sigma_\alpha \frac{\partial f}{\partial \alpha}')$ provided that $\frac{\partial f}{\partial \alpha} \neq 0$. Since impulse responses are differentiable functions of the VAR parameters and of the covariance matrix, their asymptotic distribution can be easily obtained.

Let $\mathsf{S} = [I, 0, \ldots, 0]$ be a $m \times mq$ selection matrix so that $y_t = \mathsf{S}\mathsf{Y}_t$ and $\mathsf{E}_t = \mathsf{S}'e_t$, consider the revision of the forecast at step τ and let

$$rev_t(\tau) = \mathsf{S}Rev_t(\tau) = \mathsf{S}[\mathsf{Y}_t(\tau) - \mathsf{Y}_{t-1}(\tau)] = \mathsf{S}[\mathsf{A}^\tau \mathsf{S}'\mathsf{E}_t] \equiv \psi_\tau e_t \tag{4.29}$$

We want the asymptotic distribution of the $m \times m$ matrix ψ_τ. Taking total differentials

$$d\psi_\tau = \mathsf{S}[Id\mathsf{A}\mathsf{A}^{\tau-1} + \mathsf{A}d\mathsf{A}\mathsf{A}^{\tau-2} + \ldots + \mathsf{A}^{\tau-1}d\mathsf{A}I]\mathsf{S}' \tag{4.30}$$

Since $\text{var}(\mathsf{Y}_{t+\tau}) = \mathsf{A}^{\tau} var(\mathsf{E}_{t+k})(\mathsf{A}^{\tau})'$, using the fact that $dZ = \begin{bmatrix} dZ_1 \\ 0 \end{bmatrix} = \mathsf{S}' dZ_1$ and result 4.1, we have that $vec(\mathsf{S}\mathsf{A}^j(d\mathsf{A})\mathsf{A}^{\tau-(j+1)}\mathsf{S}') = vec(\mathsf{S}\mathsf{A}^j(\mathsf{S}' d\mathsf{A}_1)\mathsf{A}^{\tau-(j+1)}\mathsf{S}') = [\mathsf{S}(\mathsf{A}^{\tau-(j+1)})' \otimes \mathsf{S}\mathsf{A}^j\mathsf{S}']vec(d\mathsf{A}_1) = [\mathsf{S}(\mathsf{A}^{\tau-(j+1)})' \otimes \psi_j]vec(d\mathsf{A}_1)$. Hence

$$\frac{vec(d\psi_\tau)}{vec(d\mathsf{A}_1)} = \sum_{j=0}^{\tau-1}[\mathsf{S}(\mathsf{A}')^{\tau-(j+1)} \otimes \psi_j] \equiv \frac{\partial vec(\psi_\tau)}{\partial vec(\mathsf{A}_1)} \tag{4.31}$$

Given (4.31), it is immediate to find the distribution of ψ_τ.

Exercise 4.30 *Derive the asymptotic distribution of ψ_τ.*

The above formulas, which use the companion form, may be computationally cumbersome when m or q are large. In these cases, the following recursive formula may be useful

$$\frac{\partial D_\tau}{\partial \alpha} = \sum_{j=1}^{max[\tau,q]} [(D'_{\tau-j} \otimes I_m)\frac{\partial A_j}{\partial \alpha} + (I_m \otimes A_j)\frac{\partial D_{\tau-j}}{\partial \alpha})] \tag{4.32}$$

Exercise 4.31 *Derive the distribution of $VD(\tau)$ for orthogonal shocks.*

Standard error bands computed with the δ-method have three problems. First, they tend to have poor properties in experimental designs featuring small scale VARs and samples of 100-120 observations. Second, the asymptotic coverage is also poor when near unit roots or near singularities are present. Third, since estimated VAR coefficients have large standard errors, impulse responses have large standard errors as well resulting, in many cases, in insignificant responses at all horizons. For these reasons, methods which employ the small sample properties of the VAR coefficients might be preferred.

Exercise 4.32 *Derive the asymptotic distribution of the τ-th term of a historical decomposition.*

• Bootstrap methods

Bootstrap standard errors, first employed in VARs by Runkle (1987), are easy to compute. Using equation (4.7) one proceeds as follows:

Algorithm 4.2

1) *Obtain $A(\ell)_{OLS}$ and $e_{t,OLS} = y_t - A(\ell)_{OLS}y_{t-1}$.*

2) *Obtain $e^l_{t,OLS}$ via bootstrap and construct $y^l_t = A(\ell)_{OLS}y^l_{t-1} + e^l_{t,OLS}, l = 1, 2, \ldots, L$.*

3) *Estimate $A(\ell)^l_{OLS}$ using data constructed in 2). Compute $D^l_j, (\tilde{D}^l_j),\ j = 1, \ldots, \tau$.*

4) *Report percentiles of the distribution of* $D_j, (\tilde{D}_j)$ *(i.e. 16-84% or 2.5-97.5%), or the simulated mean and the standard deviation of* $D_j, (\tilde{D}_j)$, $j = 1, \ldots, \tau$.

Algorithm 4.2 is easily modified to produce confidence bands for other statistics.

Example 4.18 *To compute standard error bands for the variance decomposition one would insert the calculation of* $VD_{i,i'}(\tau)^l$ *as suggested in (4.27) after step 3) of algorithm 4.2.* $VD_{i,i'}(\tau)^l$ *is the percentage of the variance of* $y_{i,t}$ *explained by* $e_{i',t}$ *at horizon* τ *in replication* l*. Then in 4) order* $VD_{i,i'}(\tau)^l$ *and report percentiles or the first two moments.*

Few remarks are in order. First, bootstrapping is appropriate when e_t is a white noise with constant variance. Therefore, the approach yields poor standard error band estimates when the lag length of the VAR is misspecified or when heteroschedasticy is present.

Since conditional heteroschedasticity is less likely to emerge with low frequency data, one possible solution is to time aggregate the data before a VAR is run and standard errors are computed.

Second, estimates of the VAR coefficients are typically biased downward in small samples. For example, in a VAR(1) with the largest root around 0.95, a downward bias of about 30 percent is to be expected even when $T = 80 - 100$. Biasedness of $A(\ell)$ is a problem because in step 2) we are generating biased y_t series. Hence, the resulting distribution is likely to be centered around an incorrect estimate of the true VAR coefficients.

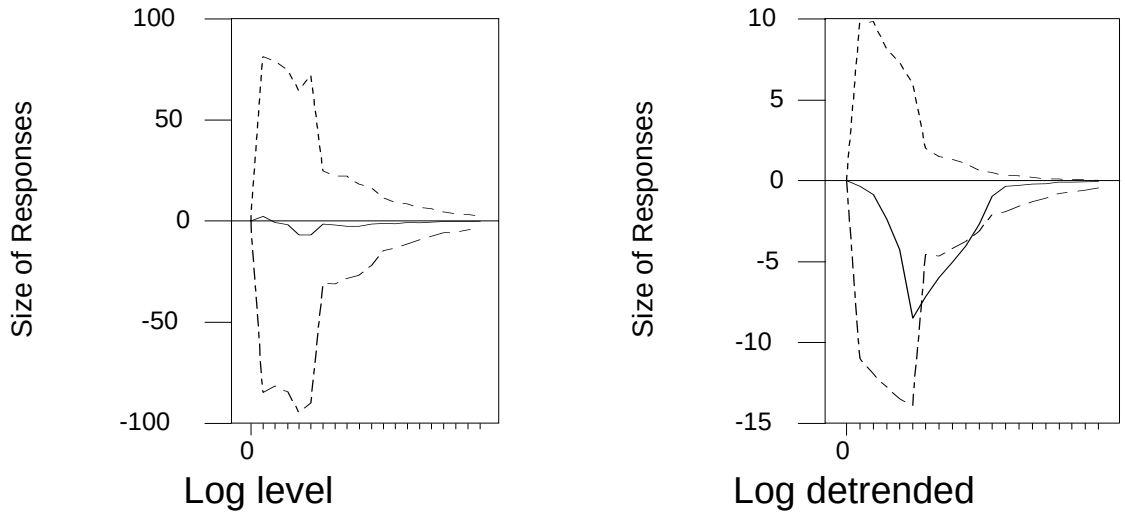


Figure 4.2: Bootstrap responses

Third, the bootstrap distribution of $D_j(\tilde{D}_j)$ is not scale invariant. In particular, units matter. This implies that standard error bands may not include, point estimates of the impulse responses. Such a problem emerges e.g. in figure 4.2, where we report one standard error bands for log output (right panel) or log linearly detrended level of output (left panel)

in response to an orthogonal price shock in a bivariate VAR(4) system together with the point estimate of the responses. Clearly, the size and the shape of the band depend on the units. Furthermore, in the right panel there are horizons where the point estimate is outside the computed standard error band.

Finally, while it is common to report the mean and construct confidence bands using numerical standard deviations (across replications), this approach is unsatisfactory since it assumes symmetric distributions. Since simulated distributions of impulse responses tend to be highly skewed when $T < 100$, we recommend the use of simulated distribution percentiles in constructing confidence intervals (i.e extract the relevant band directly from the ordered replications at each horizon).

To solve the biasedness and the lack of scale invariance, Kilian (1998) has suggested a bootstrap-after-the-bootstrap procedure. The approach can be summarized as follows:

Algorithm 4.3

1) *Given* $A(\ell)_{OLS}$*, obtain* $e^l_{t,OLS}$ *and construct* $y^l_t = A(\ell)_{OLS} y^l_{t-1} + e^l_{t,OLS}, l = 1, 2, \ldots, L.$

2) *Estimate* $A(\ell)^l_{OLS}$ *for each* l*. If the bias is approximately constant in a neighbor of* $A(\ell)_{OLS}$*,* $Bias(\ell) = E[A(\ell)_{OLS} - A(\ell)] \approx E[A(\ell)^l_{OLS} - A(\ell)_{OLS}]$*.*

3) *Calculate the largest root of the system. If it is greater or equal than one, set* $\tilde{A}(\ell) = A(\ell)_{OLS}$ *- here the bias is irrelevant since estimates are superconsistent. Otherwise set* $\tilde{A}(\ell) = A(\ell)_{OLS} - Bias(\ell)_{OLS}$*, where* $Bias(\ell)_{OLS} = \frac{1}{L}\sum_{l=1}^{L}[A(\ell)^l_{OLS} - A(\ell)_{OLS}]$*.*

4) *Repeat 1)-3) of algorithm 4.2,* L_1 *times using* $\tilde{A}(\ell)$ *in place of* $A(\ell)_{OLS}$*.*

Kilian shows that the procedure of eliminating the bias, assuming that it is constant in a neighborhood of $A(\ell)_{OLS}$, has an asymptotic justification and that the bias correction becomes negligible asymptotically. It also shows that such an approach has a better small sample coverage properties than a simple bootstrap. However, when the bias is not constant in the neighborhood of $A(\ell)_{OLS}$, the properties of the bands produced from algorithm 4.3 may still be poor.

• Monte Carlo methods

Monte Carlo methods will be described in details in the last three chapters of this book. Here we describe a simple approach which allows the computation of standard error bands using the simultaneous equation representation of an unrestricted VAR(q).

As mentioned, the likelihood function of a VAR(q), $\mathcal{L}(\alpha, \Sigma_e|y_t)$, can be decomposed into a Normal portion for α, conditional on Σ_e, and a Wishart portion for Σ_e^{-1}. Assuming that no prior information for α, Σ_e is available, i.e. $g(\alpha, \Sigma_e) \propto |\Sigma_e|^{\frac{-(m+1)}{2}}$, the posterior distribution (which is proportional to the product of the likelihood and the prior) will have a form which is identical to the likelihood. Furthermore, the posterior for (α, Σ_e) will be proportional to the product of the posterior of $(\alpha|\Sigma_e, y_t)$ and of $(\Sigma_e|y_t)$. As detailed in chapter 10, the posterior for Σ_e^{-1} has a Wishart form with $T - mq$ degrees of freedom .

The posterior of $(\alpha|\Sigma_e,\ y_t)$ is normal centered at α_{OLS} with variance equals to $var(\alpha_{OLS})$. Hence, standard error bands for impulse responses can be constructed as follows:

Algorithm 4.4

1) *Generate $T-mq$ iid draws for e_t^{-1} from a $\mathsf{N}(0,(\mathsf{Y}-\mathsf{X}\mathsf{A}_{OLS})'(\mathsf{Y}-\mathsf{X}A_{OLS}))$. Compute $\Sigma_e^l=(\frac{1}{T-mq}\sum_{t=1}^{T-mq}(e_t^{-1}-\frac{1}{T-mq}\sum_{t=1}^{T-mq}e_t^{-1})^2)^{-1}$.*

2) *Draw $\alpha^l=\alpha_{OLS}+\epsilon_t^l$, where $\epsilon_t^l\sim\mathsf{N}(0,\Sigma_e^l)$. Compute $D_j^l(\tilde{D}_j^l)$, $j=1,\dots\tau$.*

3) *Repeat 1)-2) L times and report percentiles.*

Three features of algorithm 4.4 are important. First, the posterior distribution is exact and conditional on the OLS estimator - which summarizes the information contained in the data. Therefore, biasedness of $A(\ell)_{OLS}$ is not an issue. Second, given the exact small sample nature of the posterior distribution, standard error bands are likely to be skewed and, possibly, leptokurtic. Therefore, bands extracted from percentiles are preferable to 1 or 2 standard error bands around mean. Third, algorithm 4.4 is appropriate only for just identified systems (both of semi-structural or of structural types). When the VAR system is overidentified, the technique described in section 3 of chapter 10 should be used.

Exercise 4.33 *Show how to use algorithm 4.4 to compute confidence bands for variance and historical decompositions.*

All three approaches we have described produce standard error bands estimates which are correlated. This is because responses at each step are correlated (see e.g. the recursive computation of impulse responses). Hence, plots connecting the points at each horizon are likely to misrepresent the true uncertainty. Sims and Zha (1999) propose a transformation which eliminates this correlation. Their approach relies on the following result.

Result 4.2 *If $\tilde{D}_1,\dots,\tilde{D}_\tau$ are normally distributed with covariance matrix $\Sigma_{\tilde{D}}$, the best coordinate system is given by the projection on the principal components of $\Sigma_{\tilde{D}}$.*

Intuitively, we need to orthogonalize the covariance matrix of the impulse responses to break down the correlation of its elements. To implement such an orthogonalization, for structural coefficients, steps 1) to 3) of algorithm 4.4 remain unchanged, but we need to add the following two steps

4) Let the $\tau\times\tau$ covariance matrix of $\tilde{D}$ be decomposed as $\mathcal{P}_{\tilde{D}}\mathcal{V}_{\tilde{D}}\mathcal{P}'_{\tilde{D}}=\Sigma_{\tilde{D}}$, where $\mathcal{V}_{\tilde{D}}=diag\{v_j\}$ $\mathcal{P}_{\tilde{D}}=col\{pp_{.,\ j}\},j=1,\dots,\tau,$ $\mathcal{P}_{\tilde{D}}\mathcal{P}'_{\tilde{D}}=I$.

5) For each (i,i') report $\tilde{D}^*(i,i')\pm\sum_{j=1}^{\tau}\varrho_j pp_{.,j}$, where $\tilde{D}^*(i,i')$ is the mean of $\tilde{D}(i,i')$ and $\varrho_j=pp_{j,.}\tilde{D}(i,i')$.

In practice, it is often sufficient to use the largest eigenvalue of $\Sigma_{\tilde{D}}$ to have a good idea of the existing uncertainty. Then standard error bands are $\tilde{D}^*(i,i') \pm pp_{.,j}\sqrt{v_{\sup}}$ (symmetric) and $[(\tilde{D}^*(i,i') - \varrho_{sup,.16}; \tilde{D}^*(i,i') + \varrho_{sup,.84}]$ (asymmetric), where $\varrho_{sup,.r}$ is the r-th percentile of ϱ_j computed using the largest eigenvalue of $\Sigma_{\tilde{D}}$ and $v_{\sup} = \sup_j v_j$

Exercise 4.34 *Show how to apply the Sims and Zha approach to orthogonalize standard error bands computed with the δ-method.*

Given that the asymptotic approach has poor small sample properties, which of the two resampling methods should one prefer? A-priori the choice is difficult: the bootstrap method does not require distributional assumptions but it requires homoschedasticity. Also, unless Kilian method is used, bands may have little meaning. The MC approach works even with heteroschedasticity but normality of the errors or a large sample are required. The question is therefore empirical. Sims and Zha (1999) show that, in specific experiments, the MC approach outperforms the bootstrap approach but not uniformly so.

4.4.5 Generalized Impulse Responses

This subsection discusses the computation of impulse responses for nonlinear structures. Since VARs with time varying coefficients fit well into this class, it is worthwhile to study how impulse responses for these models can be constructed. The discussion here is basic; more details are in Gallant, Tauchen and Rossi (1993) and Koop, Pesaran and Potter (1995).

In linear models impulse responses do not depend on the sign or the size of shocks nor on their history. This simplifies the computations but prevents researchers from studying interesting economic questions such as: do shocks which occur in a recession produce different dynamics than those in an expansion? Are large shocks different than small ones? In nonlinear models, responses do depend on the sign, the size and the history of the shocks up to the point where they are computed.

Let $\mathcal{F}_{t-1}$ be the history of y_{t-1} up to $t-1$. In general, $y_{t+\tau}$ depends on $\mathcal{F}_{t-1}$, the parameters α of the model and the innovations $e_{t+j}, j = 0, \ldots, \tau$. Let $Rev(\tau, \mathcal{F}_{t-1}, \alpha, e^*) = E(y_{t+\tau}|\alpha,\ \mathcal{F}_{t-1}, e_t = e^*, e_{t+j} = 0, j \geq 1) - E(y_{t+\tau}|\alpha, \mathcal{F}_{t-1}, e_{t+j} = 0, j \geq 0)$.

Example 4.19 *Consider $y_t = Ay_{t-1} + e_t$, let $\tau = 2$ and assume $|A| < 1$. Then $E(y_{t+2}|A, \mathcal{F}_{t-1}, e_{t+j} = 0,\ j \geq 0) = A^3 y_{t-1}$ and $E(y_{t+2}|A, \mathcal{F}_{t-1}, e_t = e^*,\ e_{t+j} = 0,\ j \geq 1) = A^3 y_{t-1} + A^2 e^*$ and $Rev_y(\tau, \mathcal{F}_{t-1}, A, e^*) = A^2 e^*$ which is independent of history and of the size of the shock (hence set $e^* = 1$ or $e^* = \sigma_e$) and symmetric in the sign of e^* (hence set $e^* > 0$).*

Exercise 4.35 *Consider the model $\Delta y_t = A\Delta y_{t-1} + e_t$; $|A| < 1$. Calculate the impulse response function at a generic τ. Show it is independent of the history and that the size of e^* scales the whole impulse response function. Consider an $ARIMA(d_1, 1, d_2)$: $D_1(\ell)\Delta y_t = D_2(\ell)e_t$. Show that $Rev(\tau, \mathcal{F}_{t-1}, D_2(\ell), D_1(\ell), e^*)$ is history and size independent.*

Example 4.20 *Consider the model $\Delta y_t = A_1 \Delta y_{t-1} + A_2 \Delta y_{t-1} \mathcal{I}_{[\Delta y_{t-1} \geq 0]} + e_t$, where $\mathcal{I}_{[\Delta y_{t-1} \geq 0]} = 1$ if $\Delta y_{t-1} \geq 0$ and zero otherwise. Let $0 < A = A_1 + A_2 < 1$. Then, for $e_t = e^*$*

$Rev(\tau, \Delta y_{t-1}, A, e^*) = \frac{1-A^{\tau+1}}{1-A} e^*$ *if* $\Delta y_{t-1} \geq 0$ *and* $Rev(\tau, \Delta y_{t-1}, A, e^*) = \frac{1-A_1^{\tau+1}}{1-A_1} e^*$ *if* $\Delta y_{t-1} < 0$. *Here* $Rev(\tau, \Delta y_{t-1}, A, e^*)$ *depends on the history of* Δy_{t-1}.

Exercise 4.36 *Consider the logistic map* $\tilde{y}_t = a\tilde{y}_{t-1}(1-\tilde{y}_{t-1}) + v_t$ *where* $0 \leq a \leq 4$. *This model can be transformed into a nonlinear AR(1) model:* $y_t = A_1 y_{t-1} - A_2 y_{t-1}^2 + e_t$ *for* $A_2 \neq 0,\ -2 \leq A_1 \leq 2,\ A_1 = 2-a,\ e_t = \frac{2-A_1}{A_2} v_t,\ y_t = \frac{A_1-1}{A_2} + \frac{2-A_1}{A_2}\tilde{y}_t$. *Simulate the impulse response function. Does the sign and the size of* e^* *matter?*

In impulse responses computed from linear models $e_{t+j} = 0,\ \forall j \geq 1$. This is inappropriate in nonlinear models since it may violate bounds for e_t. In exercise 4.36 the bounds occur because the logistic map is unstable if y_{t-1} passes a threshold. These bounds depend on the realizations of $v_{t-\tau}$ and therefore vary over time. Also, when parameters are estimated, we either need to condition on a particular α (e.g. α_{OLS}) or integrate α out to compute forecast revisions. Generalized impulse (GI) responses are designed to meet all these requirements: in fact we condition on the size, the sign, the history of the shocks and, if required, on a particular estimate of α and integrate out all future shocks.

Definition 4.5 *Generalized impulse responses conditional on a shock* e_t, *a history* $\mathcal{F}_{t-1}$ *and a vector* α *are* $GI_y(\tau, \mathcal{F}_{t-1}, \alpha, e_t) = E(y_{t+\tau}|\alpha, e_t,\ \mathcal{F}_{t-1}) - E(y_{t+\tau}|\alpha,\ \mathcal{F}_{t-1})$.

Responses produced by definition 4.5 have three important properties. First $E(GI_y) = 0$. Second, $E(GI_y|\mathcal{F}_{t-1}) = 0$. Third, $E(GI_y|e_t) = E(y_{t+\tau}|e_t) - E(y_{t+\tau})$.

Example 4.21 *Three interesting cases where definition 4.5 is useful are the following:*

- *(Impulse responses in recession): GI conditional only on a history* $\mathcal{F}_{t-1}$ *in a region:* $GI_y(\tau, \mathcal{F}_{t-1} \in \mathcal{F}_1, \alpha, e_t) = E(y_{t+\tau}|\alpha,\ \mathcal{F}_{t-1} \in \mathcal{F}_1, e_t) - E(y_{t+\tau}|\alpha,\ \mathcal{F}_{t-1} \in \mathcal{F}_1)$.
- *(Impulse responses on average over histories): We have two options. GI conditional only on* α*:* $GI_y(\tau, \alpha, e_t) = E(y_{t+\tau}|\alpha, e_t) - E(y_{t+\tau}|\alpha)$ *and GI unconditional on* α *:* $GI_y(\tau, e_t) = E(y_{t+\tau}|e_t) - E(y_{t+\tau})$.
- *(Impulse responses if oil prices go above 40 dollars a barrel) GI conditional on a shock in a region:* $GI_y(\tau, \mathcal{F}_{t-1}, \alpha, e_t) = E(y_{t+\tau}|\mathcal{F}_{t-1}, \alpha, e_t \in E_1) - E(y_{t+\tau}|\mathcal{F}_{t-1}, \alpha)$

Definition 4.5 conditions on a particular value of α. In some situations we may want to treat parameters as random variables. This is important in applications where symmetric shocks may have asymmetric impact on y_t depending on the value of α. Alternatively, we may want to average α out of GI. As an alternative to definition 4.5 one could use:

Definition 4.6 *Generalized impulse responses, conditional on a shock* e_t *and a history* $\mathcal{F}_{t-1}$, *are* $GI_y(\tau, \mathcal{F}_{t-1}, e_t) = E(y_{t+\tau}|\mathcal{F}_{t-1}, e_t) - E(y_{t+\tau}|\mathcal{F}_{t-1})$.

Exercise 4.37 *Extend definitions 4.5-4.6 to condition on the size and the sign of* e_t.

In practice, GI are computed numerically using Monte Carlo methods. We show how to do this conditional a on history and a set of parameters in the next algorithm.

Algorithm 4.5

1) *Fix* $y_{t-1} = \hat{y}_{t-1}, \ldots, y_{t-\tau} = \hat{y}_{t-\tau};\ \alpha = \hat{\alpha}$.

2) *Draw* $e^l_{t+j}, j = 0, 1, \ldots,$ *from* $\mathsf{N}(0, \Sigma_e)$, $l = 1, \ldots L$ *and compute* $GI^l = (y^l_{t+\tau} | \hat{y}_{t-1}, \ldots \hat{y}_{t-\tau}, \hat{\alpha},\ e_t,\ e^l_{t+j},\ j > 1) - (y^l_{t+\tau} | \hat{y}_{t-1}, \ldots \hat{y}_{t-\tau}, \hat{\alpha},\ e_t = 0,\ e^l_{t+j},\ j > 1)$.

3) *Compute* $GI = \frac{1}{L}\sum_{l=1}^{L} GI^l$, $E(GI^l - GI)^2$ *and/or the percentiles of the distribution.*

Note that in algorithm 4.5 the history $(y_{t-1}, \ldots, y_{t-\tau})$ could be a recession or expansion and $\hat{\alpha}$ an OLS or a posterior estimator. In practice, when the model is multivariate we need to orthogonalize the shocks so as to be able to measure the effect of a shock. When e_t is normal, its response to a shock in $e_{i't}$ is $E(e_t | e_{i't} = e^*_{i'}) = E(e_t e_{i't})\sigma^{-2}_{i''} e^*_{i'}$ where $\sigma^2_i = E(e_{i't})^2$ and this can be inserted in step 2) of algorithm 4.5 to compute GI. For example, for a linear VAR $GI(\tau, \mathcal{F}_{t-1}, e_{it}) = (\frac{A_\tau E(e_t, e_{i't})}{\sigma_i})\frac{e^*}{\sigma_i}$ and the generalized impulse of variable i equals $\mathsf{S}_i GI(\tau, \mathcal{F}_{t-1}, e_{it})$ where S_i is a selection vector with one in the i-th position and zero everywhere else. Here the term $\frac{e^*}{\sigma_i}$ is a scale factor and the first term measures the effect of a one standard error shock in the $i' - th$ variable. Note also, that $(\frac{A_\tau E(e_t, e_{i't})}{\sigma_i})$ corresponds to the effect obtained when the variables are assumed to have a Wold causal chain. Hence meaningful interpretations are possible only if the orthogonalization is derived from relevant economic restrictions.

Exercise 4.38 *Describe a Monte Carlo method to compute GI without conditioning on a particular history or a particular* α.

Example 4.22 *Consider the model* $\Delta y_t = A_1 \Delta y_{t-1} + A_2 \Delta y_{t-1} \mathcal{I}_{[\Delta y_{t-1} \geq 0]} + e_t$, *where* $\mathcal{I}_{[\Delta y_{t-1} \geq 0]}$ *is an indicator function. Then:*
• *GI responses allowing for randomness in* e_t *can be computed by fixing* $y_{t-1},\ A_1, A_2$ *and drawing* $e^l_{t+j},\ j \geq 0,\ l = 1, \ldots L$.
• *GI responses allowing for randomness in history can be computed fixing* $e_{t+j}\ j \geq 0, A_1, A_2$ *and drawing* y^l_{t-1}.
• *GI responses allowing for randomness in the parameters can be computed fixing* $y_{t-1},\ e_{t+j},\ j \geq 0$ *and drawing* $A^l_1,\ A^l_2$ *from some distribution (e.g. the asymptotic one).*
• *GI responses allowing for randomness in the size of* e_t *can be computed fixing* $y_{t-1},\ A_1, A_2,$ $e_{t+j}\ j > 1$ *and keeping those* e^l_t *that satisfy* $e^l_t \geq e^*$ *or* $e^l_t < e^*$. *If the process is multivariate apply the above to e.g.* e_{1t}, *after averaging over draws of* $(e_{2t}, \ldots, e_{mt})$.

Exercise 4.39 *Consider a bivariate model with inflation* π *and unemployment* $UN,\ y_t = A_1 y_{t-1} + A_2 y_{t-1} \mathcal{I}_{[\pi \geq 0]} + e_t$ *where* $\mathcal{I}_{[\pi > 0]}$ *is an indicator function. Calculate GI at steps 1 to 3 for an orthogonal shock in* π *when* $\pi \geq 0$ *and when* $\pi < 0$. *Does the size of* e_t *matter?*

Exercise 4.40 *Consider a switching bivariate AR(1) model with money and output:*

$$\Delta y_t = \begin{cases} \alpha_{01} + \alpha_{11}\Delta y_{t-1} + e_{1t} & \text{if} \quad \Delta y_{t-1} \leq \Delta\bar{y}, \quad e_{1t} \sim \mathsf{N}(0, \sigma_1^2) \\ \alpha_{02} + \alpha_{12}\Delta y_{t-1} + e_{2t} & \text{if} \quad \Delta y_{t-1} > \Delta\bar{y}, \quad e_{2t} \sim \mathsf{N}(0, \sigma_2^2) \end{cases}$$

Fix the size of the shock and the parameters and compute GI as a function of history. Fix the size of shocks and the history and compute GI as function of the parameters.

We defer further discussion on the computation of impulse responses for a particular type of non-linear model to chapter 10.

4.5 Identification

So far in this chapter, economic theory has played no role. Projections methods are used to derive the Wold theorem; statistical and numerical analysis are used to estimate the parameters and the distributions of interesting functions of the parameters. Since VARs are reduced form models it is impossible to structurally interpret the dynamics induced by their disturbances unless economic theory comes into play. As seen in chapter 2, Markovian DSGE models when approximated linearly or log linearly around the steady state typically deliver VAR(1) solutions. The reduced form parameters are complicated functions of the structural ones and the resulting set of extensive cross equations restrictions could be used to disentangle the latters if one is willing to take the model seriously as the process generating the data. When doubts about the quality of the model exists, one can still conduct inference as long as a subset of the model restrictions are credible or uncontroversial. Typical restrictions employed in the literature include constraints on the short run or long run impact of certain shocks on variables or informational delays (e.g. output is not contemporaneously observed by Central Banks when deciding interest rates). As we will argue later on, these restrictions are rarely produced by DSGE models. Restrictions involving lag responses or the dynamics are generally ignored being perceived as non robust or controversial.

To conduct structural analyses, one therefore starts from an unrestricted VAR(q) where all variables appear with the same lags in each equation, estimates the parameters of the VAR by OLS, imposes a minimal set of "structural" restrictions, possibly consistent with a variety of behavioral theories, and constructs impulse responses, historical decomposition, etc. to structural shocks. In this sense, VARs are at the antipodes of maximum likelihood or generalized method of moments approaches: the majority of the theoretical restrictions are disregarded; there is no interest in estimating preference and technology parameters; and only a structural interpretation of the shocks is sought.

We first examine identification in stationary and non-stationary VAR using zero-type (or constant-type) restrictions. Afterward, we discuss identification via sign restrictions.

4.5.1 Stationary VARs

Let the reduced form VAR be

$$y_t = A(\ell)y_{t-1} + e_t \quad e_t \sim \ iid\ (0, \Sigma_e) \tag{4.33}$$

We assume that associated with (4.33) there is a structural model of the form

$$y_t = \mathcal{A}(\ell)y_{t-1} + \mathcal{A}_0\epsilon_t \quad \epsilon_t \sim \ iid\ (0, \Sigma_\epsilon = diag\{\sigma^2_{\epsilon_i}\}) \tag{4.34}$$

Equation (4.34) generically defines a class of models but it is easy to show that it is non-empty. For example, many of the log-linearized DSGE models of chapter 2, produce solutions like (4.34) with $\mathcal{A}(\ell) = \mathcal{A}(\theta)$ and $\mathcal{A}_0 = \mathcal{A}_0(\theta)$ where θ are structural parameters. Matching contemporaneous coefficients in (4.33) and (4.34) implies $e_t = \mathcal{A}_0\epsilon_t$ or

$$\mathcal{A}_0\Sigma_\epsilon\mathcal{A}_0' = \Sigma_e \tag{4.35}$$

To compute responses to structural shocks we can proceed in two steps. First, we can estimate $A(\ell)$ and Σ_e from (4.33) using the techniques described in section 3. Second, from (4.35) and given identification restrictions, we estimate Σ_ϵ, free parameters of $\mathcal{A}_0$ and the structural dynamics $\mathcal{A}(\ell)$. This two-step approach resembles the indirect least square (ILS) technique used in a system of (static) structural equations (see Hamilton (1994, p. 244)). The main difference lies in the fact that here restrictions are imposed on the covariance matrix of reduced form residuals and not on the lags of the VAR or on the exogenous variables. This is convenient: had we imposed restrictions on the lags of the VAR, joint estimation of $A(\ell)$, Σ_e, Σ_ϵ and of free parameters of $\mathcal{A}_0$ would be required.

As in simultaneous equation systems there are necessary and sufficient conditions that need to be satisfied for identification. An order condition can be calculated as follows. On the left hand side of (4.35) there are m^2 free parameters, while given the symmetry of Σ_e, the right hand side has only $(m(m+1)/2)$ free parameters. Hence, to go from reduced form to structural shocks we need, at least, $m(m-1)/2$ restrictions (with more restrictions structural shocks are overidentified).

Example 4.23 *Consider a trivariate model with hours, productivity, and interest rates. Suppose that $\mathcal{A}_0$ is lower triangular, that is, suppose that shocks to hours enter contemporaneously in the productivity and interest rate equations and that productivity shocks enter only contemporaneously in the interest rate equation. This obtains, e.g. if interest rate shocks take time to produce effects and if hours are predetermined with respect to productivity. If structural shocks are independent, $\mathcal{A}_0$ has $m(m-1)/2 = 3$ zeros restrictions. Hence, the order condition is satisfied.*

Example 4.24 *Consider VAR with includes output, prices, nominal interest rates and money, $y_t = [GDP_t, p_t, i_t, M_t]$. Suppose that a class of models suggests that output contemporaneously reacts only to its own shocks; that prices respond contemporaneously to output and money shocks; that interest rates respond contemporaneously only to money shocks, while money contemporaneously responds to all shocks. Then $\mathcal{A}_0 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ a^0_{12} & 1 & 0 & a^0_{22} \\ 0 & 0 & 1 & a^0_{31} \\ a^0_{41} & a^0_{42} & a^0_{43} & 1 \end{bmatrix}$. Since there are six (zero) restrictions, structural shocks are identifiable.*

Exercise 4.41 *Suppose we have extraneous information which allows us to pin down some of the parameters of $\mathcal{A}_0$. For example, suppose in a trivariate system with output, hours and taxes, we can obtain estimates of the elasticity of hours with respect to taxes. How many restrictions do you need to identify the shocks? Does it make a difference if zero or constant restriction is used?*

Exercise 4.42 *Specify and estimate a bivariate VAR using Euro area GDP and M3 growth. Using the restriction that output growth is not contemporaneously affected by money growth shocks, trace out impulse responses and evaluate the claim that money has no medium-long run effect on output. Repeat the exercise assuming that the contemporaneous effect of money growth on output growth is in the interval [-0.5, 1,5] (do this in increments of 0.1 each). What can you say about the medium-long run effect of money growth on output growth in general?*

There is one additional (rank) condition one should typically check: i.e. $rank(\Sigma_e) = rank(\mathcal{A}_0\Sigma_\epsilon\mathcal{A}_0)$ (see Hamilton (1994) for a formal derivation). Intuitively, this restriction rules out that any column of $\mathcal{A}_0$ can be expressed as linear combination of the others. While the rank condition is typically important in large scale SES, it is almost automatically satisfied in small scale VAR identified with economic theory restrictions. When other types of restrictions are employed, the condition should be always checked.

Rank and order conditions are only valid for "local identification". That is, the system may not be identified even though $m(m-1)/2$ restrictions are imposed. This requires experimenting with different initial conditions when estimating the parameters of $\mathcal{A}_0$.

Example 4.25 *Suppose $\Sigma_\epsilon = I$ and that $\mathcal{A}_0^1 = \begin{bmatrix} 1 & 4 \\ 0 & 3 \end{bmatrix}$. It is immediate to verify that the likelihood obtained with these two matrices and any positive definite Σ_e is equivalent to the one obtained with the same Σ_ϵ and Σ_e and $\mathcal{A}_0^2 = \begin{bmatrix} 5 & 0.8 \\ 0 & 0.6 \end{bmatrix}$. Clearly the two decompositions have different economic interpretations. Depending on the initial conditions, the maximum can be reached at $\mathcal{A}_0^2$ or $\mathcal{A}_0^1$.*

To estimate the free parameters in (4.35) one typically has two options. The first is to write down the likelihood function of (4.35) (conditional on Σ_e), that is

$$\ln \mathcal{L} = 2 \ln |\mathcal{A}_0| + \ln |\Sigma_\epsilon| + trace(\Sigma_\epsilon^{-1}\mathcal{A}_0^{-1}\Sigma_e\mathcal{A}_0^{-1\prime}) \tag{4.36}$$

Maximizing (4.36) with respect to Σ_ϵ and concentrating it out we obtain $2 \ln |\mathcal{A}_0| + \sum_{i=1}^{m} \ln(\mathcal{A}_0^{-1}\Sigma_e\mathcal{A}_0^{-1\prime})_{ii}$. An estimate of the parameters can be found maximizing this expression with respect to the free entries of $\mathcal{A}_0$. Since the concentrated likelihood is nonstandard, maximization is typically difficult. Therefore, it is advisable to get some estimates with a simple method (e.g. a simplex algorithm) and then use these as initial conditions in other algorithms (see chapter 6) to find a global maximum.

A likelihood approach is general and works with both just-identified and overidentified systems. For a just identified system one could also use instrumental variables, as suggested, e.g. by Shapiro and Watson (1988). We describe in a example how this can be done.

Example 4.26 *Consider a bivariate VAR model with inflation and unemployment. Suppose that theory tells us that the structural system (4.34) is*

$$\begin{bmatrix} \pi_t \\ UN_t \end{bmatrix} = \begin{bmatrix} \mathcal{A}_{11}(\ell) & \mathcal{A}_{12}(\ell) \\ \mathcal{A}_{21}(\ell) & \mathcal{A}_{22}(\ell) \end{bmatrix} \begin{bmatrix} \pi_{t-1} \\ UN_{t-1} \end{bmatrix} + \begin{bmatrix} 1 & 0 \\ \alpha_{01} & 1 \end{bmatrix} \begin{bmatrix} \epsilon_{1t} \\ \epsilon_{2t} \end{bmatrix}$$

Since $\epsilon_{1t} = e_{1t}$ is predetermined with respect to ϵ_{2t} it can used as an instrument to estimate α_{01}. Therefore choosing as a vector of instruments $z_t = [e_{1t}, e_{1t-1}, \ldots, e_{2t-1}, \ldots]$ joint estimates of the α and $\mathcal{A}(\ell)$ can be obtained, applying the IV techniques described in chapter 5.

4.5.2 Nonstationary VARs

The identification process in non-stationary VAR models is similar but additional identification restrictions are available. Furthermore, the presence of cointegration constraints may change the nature of the order condition.

Let the MA representations of the VAR and of the structure be

$$\Delta y_t = D(\ell)e_t = D(1)e_t + D^*(\ell)\Delta e_t \tag{4.37}$$

$$\Delta y_t = \mathcal{D}(\ell)\mathcal{A}_0\epsilon_t = \mathcal{D}(\ell)(1)\mathcal{A}_0\epsilon_t + \mathcal{D}^*(\ell)\mathcal{A}_0\Delta\epsilon_t \tag{4.38}$$

where $D^*(\ell) \equiv \frac{D(\ell)-D(1)}{1-\ell}$, $\mathcal{D}^*(\ell) \equiv \frac{\mathcal{D}(\ell)-\mathcal{D}(1)}{1-\ell}$ and $\Delta = (1-\ell)$.

In (4.37)-(4.38) we have rewritten the system in two ways: the first is a standard MA; the second exploits the multivariate BN decomposition (see chapter 3). Matching coefficients we have $\mathcal{D}(\ell)\mathcal{A}_0\epsilon_t = D(\ell)e_t$. Separating permanent and transitory components and using in the latter case only contemporaneous restrictions we have

$$\mathcal{D}(1)\mathcal{A}_0\epsilon_t = D(1)e_t \tag{4.39}$$

$$\mathcal{A}_0\Delta\epsilon_t = \Delta e_t \tag{4.40}$$

When y_t is stationary, $\mathcal{D}(1) = D(1) = 0$, (4.39) is vacuous and only (4.40) is available. However, if y_t is integrated the restrictions linking the permanent components of the reduced and of the structural form could also be used for identification. (4.39) is the basis, e.g., for the Blanchard and Quah's decomposition discussed in chapter 3. To obtain estimates of structural parameters we need the same order and rank restrictions. However, the $m(m-1)/2$ constraints could be placed either on (4.39) or (4.40) or both. In this latter case iterative approaches are needed to estimate the free parameters of $\mathcal{A}_0$ and the structural shocks ϵ_t.

Example 4.27 *In a bivariate VAR system imposing (4.39) is simple since only one restriction is needed. Suppose that $\mathcal{D}(1)^{12} = 0$ (i.e. ϵ_{2t} has no long run effect on y_{1t}). If $\Sigma_\epsilon = I$ the three elements of $\mathcal{D}(1)\mathcal{A}_0\Sigma_\epsilon\mathcal{A}_0'\mathcal{D}(1)'$ can be obtained from the Choleski factorization of $D(1)\Sigma_e D(1)'$.*

Exercise 4.43 *Consider the model of example 4.24 and assume that all variables are integrated. Suppose we impose the same 6 restrictions via the long run multipliers $\mathcal{D}(1)\mathcal{A}_0$. Describe how to undertake maximum likelihood estimation of the free parameters.*

Exercise 4.44 *(Gali) Consider a structural model of the form*

$$\begin{aligned} y_t &= \alpha_0 + \epsilon_t^S - \alpha_1(i_t - E_t\Delta p_{t+1}) + \epsilon_t^{IS} && (4.41)\\ M_t - p_t &= \alpha_2 y_t - \alpha_3 i_t + \epsilon_t^{MD} && (4.42)\\ \Delta M_t &= \epsilon_t^{MS} && (4.43)\\ \Delta p_t &= \Delta p_{t-1} + \alpha_4(y_t - \epsilon_t^S) && (4.44) \end{aligned}$$

where ϵ_t^S is a supply shock; ϵ_t^{IS} is an IS shock; ϵ_t^{MS} is a money supply shock and ϵ_t^{MD} is a money demand shock, GDP_t is output, P_t prices, i_t the nominal interest rate and M_t money. Identify these shocks from a VAR with $(\Delta GDP_t,\ \Delta i_t,\ i_t - \Delta p_t,\ \Delta M_t - \Delta p_t)$ using Euro area data and the following restrictions: (i) only supply shocks have long run effects on output, (ii) money demand and money supply shocks have no contemporaneous effects on ΔGDP, (iii) money demand shocks have no contemporaneous effect on the real interest rate. Trace out the effects of a money supply shock on interest rates and output.

When some of the variables of the system are cointegrated, the number of permanent structural shocks is lower than m. Therefore, if long run restrictions are used, one only needs $(m-m_1)(m-m_1-1)/2$ constraints to identify all m shocks where m_1 is the number of common trends (rank of $\mathcal{D}(1) = m - m_1$).

Example 4.28 *As shown in exercise 3.4 of chapter 3, a RBC model driven by integrated technology shocks implies that all variables are integrated but $\frac{C_t}{GDP_t}$ and $\frac{Inv_t}{GDP_t}$ are stationary. Consider a trivariate VAR with $\Delta gdp_t, c_t - gdp_t, inv_t - gdp_t$ where lower case letters indicate logarithms of the variables. Since the system has two cointegrating vectors, there is one permanent shocks and two transitory ones and $(1,1,1)'\epsilon_t = D(1)e_t$ identifies the permanent shock. If all structural shocks are orthogonal we need one extra restriction to identify the two transitory disturbances - for example, we could assume a Choleski structure.*

Exercise 4.45 *(Shapiro and Watson) Consider a bivariate system $\Delta y_t = D(\ell)e_t$ where $e_t \sim (0, \Sigma_e)$ and let the structural model be $\Delta y_t = \mathcal{D}(\ell)\epsilon_t$ where $\epsilon_t \sim (0, I)$ and $\mathcal{D}(1)$ is lower triangular. Show that $D(1) = \mathcal{D}(1)\Sigma_e^{0.5}$ is lower triangular. Show that to estimate $\mathcal{D}(1)$ and $\mathcal{D}_0$ one could normalize the system $\Delta y_t^* = \Sigma_e^{-0.5}\Delta y_t$ and run a regression of $\Delta y*_{1t}$ on q lags of Δy_{1t}^* and the current and $q-1$ lags of Δy_{2t}^* and a regression of $\Delta y*_{2t}$ on q lags of Δy_{1t}^* and Δy_{2t}^*, instrumenting current values with $\Delta y_{t-j}^*, j = 1, 2, \ldots$.*

4.5.3 Alternative identification schemes

The identification of structural shocks is, in general, a highly controversial enterprise because researchers imposing different identifying assumptions may reach different conclusions about

interesting economic questions (e.g. the sources of business cycle fluctuations). However, an embarrassing uniformity has emerged over the last 10 years since identifying restrictions have become largely conventional and unrelated to the class of DSGE models described in section 2. Criticisms to the nature of identification process have repeatedly appeared in the literature. For example, Cooley and LeRoy (1985) criticize Choleski decompositions because contemporaneous recursive structures are hard to obtain in general equilibrium models. Faust and Leeper (1997) argue that long run restrictions are unsatisfactory as they may exclude structures which generate perfectly reasonable short run dynamics but fail to satisfy long run constraints by infinitesimal amounts. Cooley and Dwyer (1998) indicate that long restrictions may also incompletely disentangle permanent and transitory disturbances. Canova and Pina (2004) show that standard DSGE models almost never provide the zero restrictions employed to identify monetary disturbances in structural VAR systems and that misspecification of the features of the underlying economy can be substantial.

Figure 4.3 shows the extent of the problem when a working capital model, similar to the one presented in chapter 2, with either a partial accommodative (PA) or a Taylor type (FB) rule for monetary policy is used to generate data and monetary shocks are identified in the VAR for simulated data either with a Choleski scheme (CEE), with variables in the order $(GDP_t, p_t, i_t, \frac{M_t}{p_t})$ or via an overidentified structure (SZ) where i_t responds only to $\frac{M_t}{p_t}$. The straight line is the response produced by the model, the dotted ones one standard error bands produced by the VAR. Note that a Choleski system correctly recognizes the policy input when a Taylor rule is used, while the overidentified model correctly characterizes the policy rule in the partial accommodative case. Misspecification is pervasive even when one correctly selects the inputs of the monetary policy rule. For example, a Choleski scheme fails to capture the persistent response of real balances to interest rate increases and produces perverse output responses (first box, first column) while a price puzzle is produced (second row, first and third boxes).

We would like to stress that the patterns presented in figure 4.3 are not obtained because the model is unrealistic or the parametrization "crazy". As shown in Canova and Pina (2004) a sticky price, sticky wage model, parametrized in a standard way, produces similar outcomes. The problem is that a large class of DSGE structures do not display the zero restrictions imposed by the two identification schemes (in particular, that output and prices have a Wold causal structure and do not respond instantaneously to policy shocks). Therefore, misspecification results even when the policy rule is correctly identified.

To produce a more solid bridge between DSGE models and VARs, a new set of identification approaches have emerged. Although justified with different arguments, the procedures of Faust (1998), Uhlig (2003) and Canova and De Nicoló (2002) have one feature in common: they do not use zero-type of restrictions. Instead, they achieve identification restricting the sign (and/or shape) of structural responses. Restrictions of this type are often used by applied researchers informally: for example, monetary shocks which do not generate a liquidity effect (e.g. opposite comovements in interest rate and money) are typically discarded and the zero restrictions reshuffled in the hope to produce the required outcome. One advantage of these approaches is to make restrictions of this type explicit.

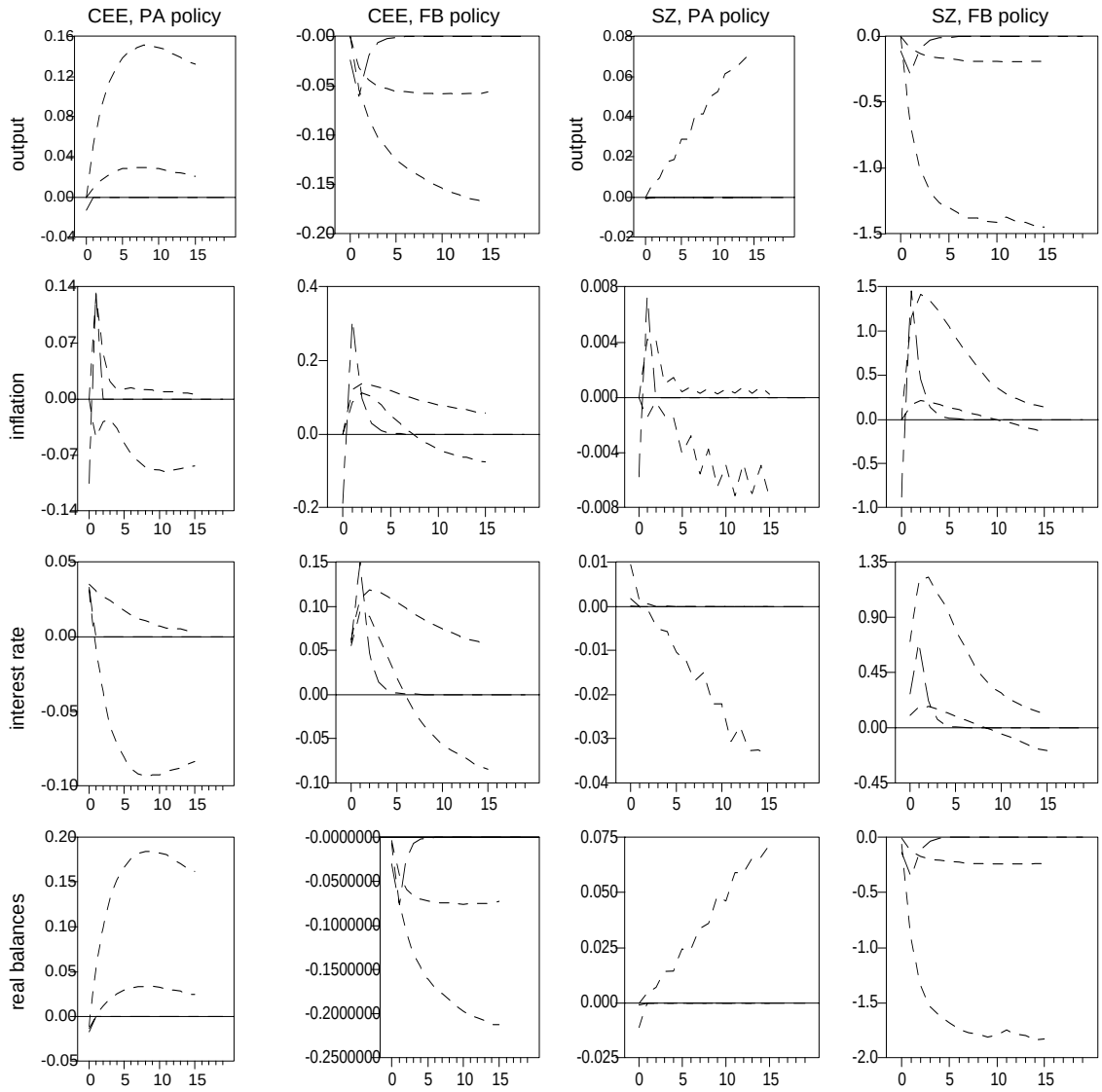


Figure 4.3: Impulse responses to monetary shocks, Working capital model

Sign restrictions are enticing. While (log)-linearized versions of DSGE models seldomly deliver the $m(m-1)/2$ set of zero restrictions needed to recover m structural shocks, they contain a large number of sign restrictions usable for identification purposes.

Example 4.29 *(Technology shocks) All RBC-type models examined in chapter 2 have the feature that positive technology disturbances increase output, consumption, and investment either instantaneously or with a short lag, while prices and interest rates decline as the aggregate supply curve shifts to the right. Therefore, such a class of models suggests that technology disturbances can be identified via the restriction that in response to positive shocks real variables increase and prices decrease, either contemporaneously or with a lag.*

Example 4.30 *(Monetary shocks) Several of the models of chapter 2 have the feature that policy driven increases in the nominal interest rates reduce real balances instantaneously and induce a fall in prices. Hence, contemporaneous (and lagged) comovements of real balances, prices and nominal interest rates can be used to identify monetary disturbances.*

The restrictions of examples 4.29 and 4.30 could be imposed on two or more variables, at one or more horizons. In other words, we can "weakly" or "strongly" identify the shocks. To maintain comparability with other structural VARs, weak forms of identification will be typically preferred. However, one should be aware that restrictions which are too weak may be unable to distinguish shocks with somewhat similar features, i.e. labor supply and technology shocks.

It is relatively complicated to impose sign restrictions directly on the coefficients of the VAR, as this requires maximum likelihood estimation of the full system under inequality constraints. However, it is relatively easy to do it ex-post on impulse responses. For example, as in Canova De Nicolo'(2002), one could estimate $A(\ell)$ and Σ_e from the data using OLS and orthogonalize the reduced form shocks using, e.g. an eigenvalue-eigenvector decomposition, $\Sigma_e = \mathcal{P}\mathcal{V}\mathcal{P}' = \tilde{\mathcal{P}}\tilde{\mathcal{P}}'$ where $\mathcal{P}$ is a matrix of eigenvectors and $\mathcal{V}$ is a diagonal matrix of eigenvalues. This decomposition does not have any economic content, but produces uncorrelated shocks without employing zero restrictions. For each of the orthogonalized shocks one can check if the identifying restrictions are satisfied. If there is one such a shock, the process terminates. If there is more than one shock satisfying the restrictions, one may want to increase the number of restrictions (either across variables or across leads and lags) until one candidate remains or take an average. Practical experience suggests that contemporaneous and/or one lag restrictions suffice to produce a unique set of shocks.

If no shock satisfies the restrictions, the non-uniqueness of the MA representation can be used to provide alternative structural shocks. In fact, for any $\mathcal{H}$ with $\mathcal{H}\mathcal{H}' = I$, $\Sigma_e = \tilde{\mathcal{P}}\tilde{\mathcal{P}}' = \tilde{\mathcal{P}}\mathcal{H}\mathcal{H}'\tilde{\mathcal{P}}'$. Hence, one can construct a new decomposition using $\tilde{\mathcal{P}}\mathcal{H}$ and examine if the shocks produce the required pattern.

The only remaining practical question is how to choose $\mathcal{H}$ and how to systematically explore the space of MA representations, which is infinite dimensional, if this is of interest. Canova and de Nicoló choose $\mathcal{H} = \mathcal{H}(\omega)$, $\omega \in (0, 2\pi)$ and search the space of $\mathcal{H}$ by varying ω on a grid. Here $\mathcal{H}$ are matrices which rotate the columns of $\mathcal{P}$ by an angle ω.

Example 4.31 *Consider a bivariate system with unemployment and inflation and suppose that a basic eigenvector-eigenvalue decomposition has not produced a shock which produced contemporaneously negative comovements in inflation and unemployment. Set* $\mathcal{H}(\omega) = \begin{bmatrix} cos(\omega) & -sin(\omega) \\ sin(\omega) & cos(\omega) \end{bmatrix}$. *Then we can trace out all possible MA representations for the bivariate system, varying* $\omega \in (0, 2\pi)$.

In larger scale systems, rotation matrices are more complex.

Exercise 4.46 *Consider a four variable VAR. How many matrices rotating two or pairs of two columns exist? How would you explore the space of rotations simultaneously flipping the first and the second column together with the third and the fourth?*

When m is of medium size, the matrix $\mathcal{H}$ has the following form

$$\mathcal{H}_{i,i'}(\omega) = \begin{pmatrix} 1 & 0 & 0 & \dots & 0 & 0 \\ 0 & 1 & 0 & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & \cos(\omega) & \dots & -\sin(\omega) & 0 \\ \vdots & \vdots & \vdots & 1 & \vdots & \vdots \\ 0 & 0 & \sin(\omega) & \dots & \cos(\omega) & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

$0 < \omega \leq 2\pi$ where the index (i, i') indicates that columns i and i' are rotated by the angle ω. Let $\mathcal{Z}(\mathcal{H}_{i,i'}(\omega)$ be the space of ortonormal rotation matrices where, given ω, each i, i' element has proability $\frac{2}{m(m-1)}$. Then the following search algorithm could be used to explore the space of identifications.

Algorithm 4.6

1) *Draw* ω^l *from* $(0, 2\pi)$. *Draw* $\mathcal{H}_{i,i'}(\omega^l)$ *from* $\mathcal{Z}(\mathcal{H}_{i,i'}(\omega^l)$.

2) *Used* $\mathcal{H}(i, i')(\omega^l)$ *to compute* ϵ_t *and* $\mathcal{A}(\ell)$. *Check whether restrictions are satisfied in response to* $\epsilon_{it}, i = 1, \dots m$. *If they are keep the draw, if they are not, drop the draw.*

3) *Repeat 1) and 2) unless L draws satisfying the restrictions are found. Report percentile response bands.*

Note that, by continuity, it is typical to find an interval (ω_1, ω_2) which produces a shock with the required characteristics. Since within this interval the dynamics produced by structural shocks are similar, one can average statistics for all the shocks in the interval or choose, say, the shock corresponding to the median point of the interval or keep all of them, as we have done in algorithm 4.6. We have already discussed what to do if more than one ϵ_{it} satisfies the restrictions for a given ω^l and $\mathcal{H}_{i,i'}(\omega^l)$. At times one may find disjoint intervals

where one or more shocks satisfy the restrictions. In this case it is a good idea to graphically inspect the outcome since responses may not be economically meaningful (for example, a shock may imply an output elasticity of 50). When visual inspection fails, increasing the number of restrictions is typically sufficient to eliminate "unreasonable" intervals.

Exercise 4.47 *Provide a Monte Carlo algorithm to construct standard error bands for structural impulse responses identified with sign restrictions which takes into account parameter uncertainty.*

Example 4.32 *Figure 4.4 presents the responses of industrial output, prices and M1 in the US in response to a monetary policy shock. In the right column are the 68% impulse response bands obtained requiring that a nominal interest rate increase must be accompanied by a liquidity effect - a contemporaneous decline in M1. In the left column are the 68% impulse response bands obtained with the Choleski system where the interest rate is assumed to contemporaneously react to industrial output and prices but not to money.*

Clearly, the standard identification has unpleasant outcomes: point estimates of money, output and prices are all positive after the shock even though the increase is not significant. With sign restrictions, output and prices significantly decline after a contractionary shock and they do so for about 5 months. Note that in both systems no measure of commodity prices is used.

4.6 Problems

While popular among applied researchers, VARs are not free of problems and a number of common pitfalls should be avoided when interpreting the results.

First, one should be aware of time aggregation problems. As Sargent and Hansen (1991), Marcet (1991) and others have shown time aggregation may make inference difficult. In fact, if agents take decisions every τ periods but an econometrician observes data only every $j\tau$, $j > 1$, the statistical model used by the econometrician (with data sampled at every $j\tau$) may have little to do with the one produced by agents' decisions. For example, the MA traced out by the econometrician is not necessarily the MA of the model sampled every j period, but a complex function of all MA coefficients from that point on to infinity.

Example 4.33 *Marcet (1991) showed that if agents' decisions are taken in continuous time, continuous and discrete time MA representations are related via* $D_j = [d \diamond v'_{-j}][v \diamond v_0]$ *where* d *is the moving average in continuous time,* $\diamond$ *indicates the convolution operator and* $v_j = d_j - b \times (d_j|D)$ *is the forecast error in predicting* d_j *using the information contained in the discrete time MA coefficients,* b *is a constant and* $j = 1, 2, \ldots, \tau$. *Hence a humped-shaped monthly response can easily be transformed into a smoothly declining quarterly response (see figure 4.5).*

One important special case obtains when agents' decisions generate a VAR(1) for the endogenous variables. In that case, the MA coefficients of, say, a quarterly model are

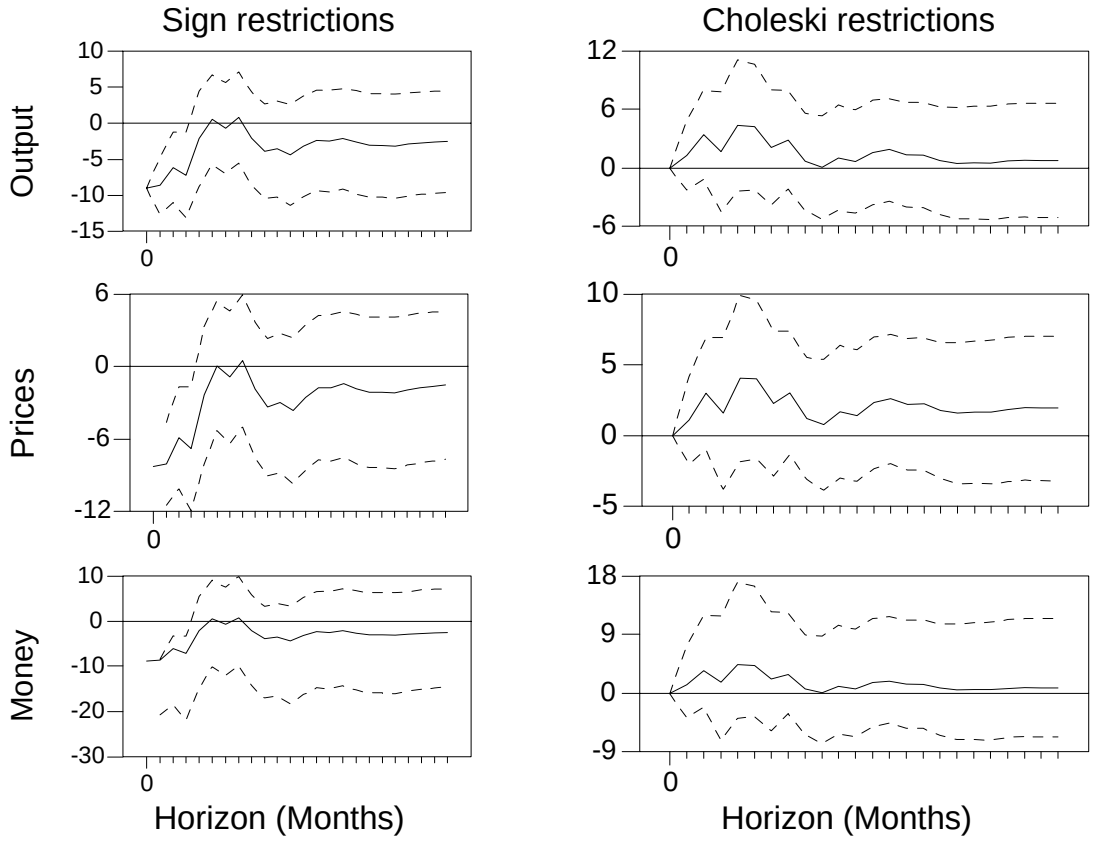


Figure 4.4: Responses to a US policy shock, 1964:1-2001:10

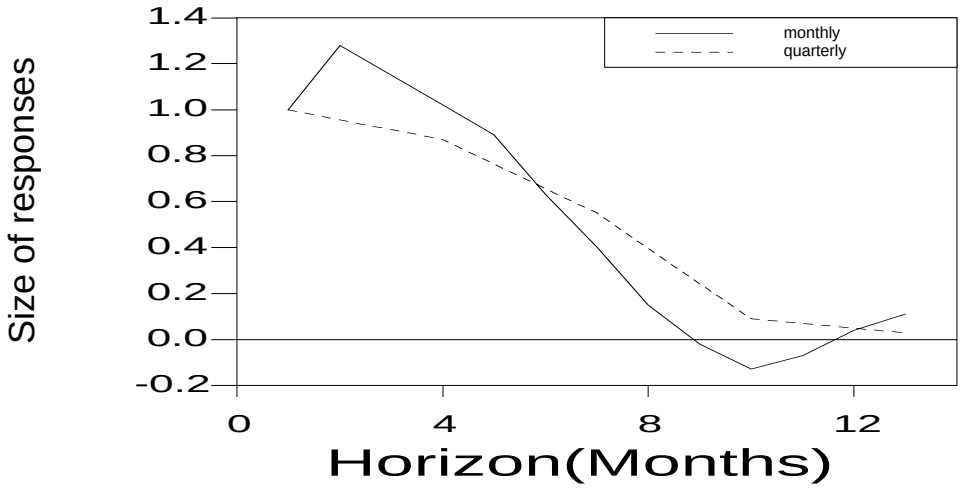


Figure 4.5: Quarterly and Monthly MA representations

the same as the quarterly sampled version of MA coefficients of a monthly model. While log-linear or quadratic approximate solutions to many DSGE models do deliver VAR(1) models, one should be aware that models with e.g. habit in consumption or quadratic costs of adjustment to investments, produce more complicated dynamics and therefore may face important aggregation problems.

Exercise 4.48 *Consider a RBC model perturbed by technology and government expenditure disturbances. Suppose that $g_t = T_t$ where T_t are lump-sum taxes and that the utility function depends on current and lagged leisure, i.e. $U(c_t, N_t, N_{t-1}) = \ln c_t + (N_t - \gamma N_{t-1})^{\varphi_n}$.*
i) Calculate the linearized decision rules after you have appropriately parametrized the model at quarterly and annual frequencies. Compare the MA coefficients of the annual model with the annual sampling of the MA of the quarterly model.
ii) Simulate consumption and output for the two specifications. Sample at annual frequencies the quarterly data and compare the autocovariance functions. Does aggregation hold?
iii) Set $\gamma = 0$ and assume that both capital and its utilization enter in the production function as in exercise 2.10 of chapter 2. Repeat steps i)- ii) and comment on the results.

Exercise 4.48 suggests that one way to detect possible aggregation problems is to run VARs at different frequencies and compare their ACF or their MA representations. If differences are detected, given the same amount of data, aggregation is likely to be a problem.

A second important problem has to with the dimensionality of the VAR. Small scale VAR models are typically preferred by applied researchers since parameter estimates are more precise (and impulse response bands are tighter) and because identification of the

structural shocks is easier. However, small scale VARs are prone to misspecification. For example, there may be important omitted variables and shocks may be confounded or misaggregated. As Braun and Mittnik (1993), Cooley and Dwyer (1998), Canova and Pina (2004) have shown, important biases may result. To illustrate the effects of omitting variables we make use of the following result:

Result 4.3 *In a bivariate VAR(q):* $\begin{bmatrix} A_{11}(\ell) & A_{12}(\ell) \\ A_{21}(\ell) & A_{22}(\ell) \end{bmatrix} \begin{bmatrix} y_{1t} \\ y_{2t} \end{bmatrix} = \begin{bmatrix} e_{1t} \\ e_{2t} \end{bmatrix}$, *the univariate representation for* y_{1t} *is* $[A_{11}(\ell) - A_{12}(\ell)A_{22}(\ell)^{-1}A_{21}(\ell)]y_{1t} = e_{1t} - A_{12}(\ell)A_{22}(\ell)^{-1}e_{2t} \equiv \upsilon_t$

Example 4.34 *Suppose the true DGP has* $m = 4$ *variables but an investigator incorrectly estimates a bivariate VAR (there are three of these models). Using result 4.3 it is immediate to see that the system with, e.g., variables 1 and 3, has errors of the form* $\begin{bmatrix} \upsilon_{1t} \\ \upsilon_{2t} \end{bmatrix} \equiv \begin{bmatrix} e_{1t} \\ e_{3t} \end{bmatrix} - \mathsf{Q}_1(\ell)\mathsf{Q}_2^{-1}(\ell) \begin{bmatrix} e_{2t} \\ e_{4t} \end{bmatrix}$ *where* $\mathsf{Q}_1(\ell) = \begin{bmatrix} A_{12}(\ell) & A_{14}(\ell) \\ A_{32}(\ell) & A_{34}(\ell) \end{bmatrix}$ $\mathsf{Q}_2(\ell) = \begin{bmatrix} A_{22}(\ell) & A_{24}(\ell) \\ A_{42}(\ell) & A_{44}(\ell) \end{bmatrix}$. *From this one can verify that:*

- *If the true system is a VAR(1), a model with* $m_1 < m$ *variables is a VAR(*∞*).*
- *If* e_t*'s are contemporaneously and serially uncorrelated,* υ_t*'s are typically contemporaneously and serially correlated.*
- *Two small scale VAR, both with* $m_1 < m$ *variables, may have different innovations.*
- υ_t *is a linear combination of current and past* e_t*. The timing of innovations is preserved if the* m_1 *included variables are Granger causally prior to the* $m - m_1$ *omitted ones (i.e. if* $\mathsf{Q}_1(\ell) = 0$*).*

Several implications one can draw from example 4.34. First, if relevant variables are omitted a long lag length is needed to whiten the residuals. While long lags do not always indicate misspecification (for example, if y_t is nearly non-stationary long lags are necessary to approximate its autocovariance function), care should be exercised in drawing inference in such models. Second, two researchers estimating small scale models with different variables may obtain different structural innovations, even if the same identification restrictions are used. Finally, innovation accounting exercises when variables are omitted may misrepresent the timing of the responses to structural shocks.

Exercise 4.49 *(Giordani) Consider a sticky price model composed of an output gap (*$gap_t = gdp_t - gdp_t^P$*) equation, a potential output (*gdp_t^P*) equation, a backward looking Phillips curve (normalized on* π_t*) and a Taylor rule of the type*

$$gap_{t+1} = a_1 gap_t - a_2(i_t - \pi_t) + \epsilon_{t+1}^{AD} \tag{4.45}$$
$$gdp_{t+1}^P = a_3 gdp_t^P + \epsilon_{t+1}^P \tag{4.46}$$
$$\pi_{t+1} = \pi_t + a_4 gdp_t^g + \epsilon_{t+1}^{CP} \tag{4.47}$$
$$i_t = a_5 \pi_t + a_6 gdp_t^g + \epsilon_{t+1}^{MP} \tag{4.48}$$

The last equation has an error term (monetary policy shock) since the central bank may not always follow the optimal solution to its minimization problem. Let $var(\epsilon^i_{t+1}) = \sigma^2_i$, $i =$ *AD, P, CP, MP and assume that the four shocks are uncorrelated with each other.*
(i) Argue that contractionary monetary policy shocks have one period lagged (negative) effects on output and two periods lagged (negative) effects on inflation. Show that monetary policy actions do not Granger cause gdp^P_t *for all* t.
(ii) Derive a VAR for $[gdp_t,\ gdp^P_t,\ \pi_t,\ i_t]$. *Display the matrix of impact coefficients.*
(iii) Derive a representation for a three variable system $[gdp_t,\ \pi_t,\ i_t]$ *(Careful: when you solve out potential output from the system the remaining variables do not follow a VAR any longer). Label the three associated shocks* $e_t = [e^{AD}_t, e^{CP}_t, e^{MP}_t]$ *and their covariance matrix* Σ_e. *Show the matrix of impact coefficients in this case.*
(iv) Show that $var(e^{AD}_t) > var(\epsilon^{AD}_t)$; $var(e^{MP}_t) > 0$ *even when* $\epsilon^{MP}_t = 0\ \forall t$ *and that* $corr(e^{MP}_t, \epsilon^P_t) < 0$. *Show that in a trivariate system, contractionary monetary policy shocks produce positive price responses (compare this with what you have in i))*
(v) Intuitively explain why the omission of potential output from the VAR causes problems.

It is worthwhile to look at omitted variable problems from another perspective. Suppose the structural MA for a partition with $m_1 < m$ variables of the true DGP is

$$y_t = D(\ell)\epsilon_t \tag{4.49}$$

where ϵ_t is an $m \times 1$ vector, so that $D(\ell)$ is $m_1 \times m$ matrix $\forall \ell$. Suppose a researcher specifies a VAR with $m_1 < m$ variables and obtains an MA of the form:

$$y_t = \tilde{D}(\ell)e_t \tag{4.50}$$

where e_t is an $m_1 \times 1$, and $\tilde{D}(\ell)$ is a $m_1 \times m_1$ matrix $\forall \ell$. Matching (4.49) and (4.50) one obtains $\tilde{D}(\ell)e_t = D(\ell)\epsilon_t$ or letting $D^\ddagger(\ell)$ be a $m_1 \times m$ matrix

$$D^\ddagger(\ell)\epsilon_t = e_t \tag{4.51}$$

As shown by Faust and Leeper (1997) (4.51) teaches us an important lesson. Assume that there are m^a shocks of one type and m^b shocks of another, $m^a + m^b = m$, and that $m_1 = 2$. Then $e_{it}, i = 1, 2$ recovers a linear combination of shocks of type $i' = a, b$ only if $D^\ddagger(\ell)$ is block diagonal and only correct current shocks if $D^\ddagger(\ell) = D^\ddagger, \forall \ell$ and block diagonal. In all other cases, true innovations are mixed up in estimated structural shocks.

Note that these problems have nothing to do with estimation or identification. Misspecification occurs because a VAR(q) is transformed in a VARMA(∞) whenever a variable is omitted and this occurs even when the MA representation of the small scale model is known.

Example 4.35 *Suppose the true structural model has* $m = 4$ *shocks, that there are two supply and two demand shocks, and that an investigator estimates a bivarate VAR. When would the two estimated structural shocks correctly aggregate shocks of the same type? Using*

(4.51) we have $\begin{bmatrix} D_{11}^{\ddagger}(\ell) & D_{12}^{\ddagger}(\ell) & D_{13}^{\ddagger}(\ell) & D_{14}^{\ddagger}(\ell) \\ D_{21}^{\ddagger}(\ell) & D_{22}^{\ddagger}(\ell) & D_{23}^{\ddagger}(\ell) & D_{24}^{\ddagger}(\ell) \end{bmatrix} \begin{bmatrix} \epsilon_{1t} \\ \epsilon_{2t} \\ \epsilon_{3t} \\ \epsilon_{4t} \end{bmatrix} = \begin{bmatrix} e_{1t} \\ e_{2t} \end{bmatrix}$. *Hence,* e_{1t} *will recover only type 1 shocks if* $D_{13}^{\ddagger}(\ell) = D_{14}^{\ddagger}(\ell) = 0$ *and* e_{2t} *will recover type 2 shocks if* $D_{21}^{\ddagger}(\ell) = D_{22}^{\ddagger}(\ell) = 0$. *Furthermore,* e_{1t} *recovers current type 1 shocks if* $D_{13}^{\ddagger}(\ell) = D_{14}^{\ddagger}(\ell) = 0$ *and* $D_{ii'}^{\ddagger}(\ell) = D_{ii'}^{\ddagger}$, $\forall \ell$.

The conditions required for correct aggregation are therefore somewhat strong. As it is shown in the next example, they are not satisfied in at least one type of DSGE model. It is likely that such a problem also appears in other models macroeconomists currently use.

Example 4.36 *We simulate data from a version of the working capital economy of example 2.14 of chapter 2 with a permanent (technology) disturbance and temporary labor supply, monetary and government expenditure shocks. Monetary policy is characterized by a Taylor rule. Using output and employment data we estimate a bivariate VAR and extract a permanent and a transitory shock where the latter is identified by the requirement that it has no long run effects on output. Table 4.3 presents the estimated coefficients of a distributed lag regression of two of the theoretical shocks on the estimated ones. In parenthesis are t-statistics. The last column presents the p-value of a F-test excluding monetary disturbances from the first equation and technological disturbances from the second. Estimated supply shocks mix both current and lagged monetary and technology disturbances while for estimated demand shocks current and lagged monetary disturbances matter but only current technology disturbances are important. This pattern is independent of the sample size.*

	Technology Shocks			*Monetary Shocks*			*P-value*
	0	*-1*	*-2*	*0*	*-1*	*-2*	
Estimated	*1.20*	*0.10*	*0.04*	*0.62*	*-0.01*	*-0.11*	
Supply Shocks	*(80.75)*	*(6.71)*	*(3.05)*	*(45.73)*	*(-0.81)*	*(-8.22)*	*0.000*
Estimated	*-0.80*	*0.007*	*0.08*	*0.92*	*-0.48*	*-0.20*	
Demand Shocks	*(-15.27)*	*(0.13)*	*(1.59)*	*(19.16)*	*(-10.03)*	*(-4.11)*	*0.000*

Table 4.3: Regressions on simulated data

Exercise 4.50 *(Cooley and Dwyer) Simulate data from a CIA model where a representative agent maximizes* $E_0 \sum_t \beta^t [a \ln c_{1t} + (1-a) \ln c_{2t} - \vartheta_N N_t]$ *subject to* $p_t c_{1t} \leq M_t + (1+i_t)B_t + T_t - B_{t-1}$ *and* $c_{1t} + c_{2t} + inv_t + \frac{M_{t+1}}{p_t} + \frac{B_{t+1}}{p_t} \leq w_t N_t + r_t K_t + \frac{M_t}{p_t} + (1+i_t)\frac{B_t}{p_t} + \frac{T_t}{p_t}$ *where* $K_{t+1} = (1-\delta)K_t + inv_t$, $y_t = \zeta_t K_t^{1-\eta} N_t^{\eta}$, $\ln \zeta_t = \rho_\zeta \ln \zeta_{t-1} + \epsilon_{1t}$, $\ln M_{t+1}^s = \ln M_t^s + \ln M_t^g$, M_t^g *is a constant and* $\rho_\zeta = 0.99$ *(you are free to choose the other parameters, but motivate your choices). Consider a bivariate system with output and hours and verify that output has a unit root but hours does not. Using the restriction that demand shocks have no long run effects on output, plot output and hours responses in theory and in the VAR. Is there any feature of the theoretical economy which is distorted?*

In section 5 we have seen that for a just identified structural model, a two-step estimation approach is equivalent to a direct 2SLS approach on the structural system. Since structural shocks depend on the identification restrictions, we may have situations where a 2SLS approach produces "good' estimators, in the sense that they nicely correlate with the structural shocks they instrument for, and situations where they are bad. Cooley and Dwyer (1998) present an example where, by changing the identifying restrictions, the correlation of the instruments with the structural shocks go from high to very low, therefore resulting in instrumental variables failures (see chapter 5). Hence, if such a problem is suspected, a maximum likelihood approach should be preferred.

Finally, we would like to mention once again that there are several economic models which generate non-Wold decompositions, see e.g. Leeper (1991), Quah (1990), Hansen and Sargent (1991). Hence examining these models with Wold decompositions is meaningless. When a researcher suspects that this is a problem Blaascke factors should be used to construct non-fundamental structural MA representations. Results do depend on the representations used. For example, Lippi and Reichlin (1993) present a non-Wold version of Blanchard and Quah (1989)'s model which gives opposite conclusions regarding the relative importance of demand and supply shocks in generating business cycle fluctuations.

Exercise 4.51 *(Quah) Consider a three equations permanent income model*

$$\begin{aligned} c_t &= rWe_t \\ We_t &= sa_t + [(1+r)^{-1}\sum_j (1+r)^{-j} E_t GDP_{t+j} \\ sa_{t+1} &= (1+r)sa_t + GDP_t - c_t \end{aligned} \tag{4.52}$$

where c_t is consumption, We_t is wealth, r is the (constant) real rate, sa_t are savings and $\Delta GDP_t = D(\ell)\epsilon_t$ is the labor income. Show that a bivariate representation for consumption and output is $\begin{bmatrix} \Delta GDP_t \\ \Delta c_t \end{bmatrix} = \begin{bmatrix} A_1(\ell) & (1-\ell)A_0(\ell) \\ A_1(\beta) & (1-\beta)A_0(\beta) \end{bmatrix} \begin{bmatrix} e_{1t} \\ e_{0t} \end{bmatrix}$ *where $\beta = (1+r)^{-1}$, e_{1t} is a permanent shock and e_{0t} a transitory shock. Find $A_1(\ell)$ and $A_0(\ell)$. Show that if $\Delta Y_t = \epsilon_t$, the representation collapses to* $\begin{bmatrix} \Delta GDP_t \\ \Delta c_t \end{bmatrix} = \begin{bmatrix} 1 & (1-\ell) \\ 1 & (1-\beta) \end{bmatrix} \begin{bmatrix} e_{1t} \\ e_{0t} \end{bmatrix}$. *Show that the determinant of the matrix vanishes at $\ell = \beta < 1$ so that the MA representation for consumption and income is non-fundamental. Show that the fundamental MA is* $\begin{bmatrix} \Delta GDP_t \\ \Delta c_t \end{bmatrix} = b(\beta)^{-1} \begin{bmatrix} (2-\beta)(1-\frac{1-\beta}{2-\beta}\ell) & (1-\beta\ell) \\ 1+(1-\beta)^2 & 0 \end{bmatrix} \begin{bmatrix} \tilde{e}_{1t} \\ \tilde{e}_{0t} \end{bmatrix}$, *var($\tilde{e}_{0t}$)=var($\tilde{e}_{1t}$) = 1.*

4.7 Validating DSGE models with VARs

VARs are extensively used to summarize those conditional and unconditional moments that "good" models should be able to replicate. Generally, informal comparisons between the models and the data are performed. At times, the model's statistics are compared with 68 or 95% bands for the statistics of the data (see e.g. Christiano Eichenbaum and Evans

(2001)). There conclusions about the quality of the model rest on whether model's statistics are inside or outside these bands for a number of variables. If parameter uncertainty is allowed for, comparison of posterior distributions is possible (see chapters 7 and 11).

However, DSGE theories can be more directly tested via VARs. For example, in Canova, Pagan and Finn (1994) theoretical cointegration restrictions coming from a RBC model driven by permanent technology shocks are imposed on a VAR and tested using standard statistical tools. Their point of view can be generalized and the applicability of their idea extended if qualitative implications, which are more robust than quantitative ones, are used to restrict the data and if restrictions are used for identification rather than for estimation.

DSGE models are misspecified in the sense that they are too simple to capture the complex probabilistic nature of the data. Hence, it may be senseless to compare their outcomes with the data: if one looks hard enough and data is abundant, statistically or economically large deviations can always be found. Both academic economists and policymakers use DSGE models to tell stories about how the economy responds to unexpected movements in exogenous variables. Hence, there may be substantial consensus in expecting output to decline after an unexpected interest rate increase but considerable uncertainty about the size of the impact and the timing of the output responses. The techniques described in chapter 5 to 7 have hard time to deal with this uncertainty. Estimation and testing with maximum likelihood requires the whole model to be the correct DGP (up to uncorrelated measurement errors), at least under the null. Generalized methods of moments and simulation estimators can be tailored to focus only on those aspects where misspecification could be smaller (e.g. the Euler equation, or the great ratios). However, estimation and validation still requires that these aspects of the model are quantitatively correct under the null. When one feels comfortable only with the qualitative implications of a model and is not willing to (quantitatively) entertain a part or the whole of it as a null hypothesis, the approach described in section 5.3 can be used to formally evaluate the fit of any model or the relative merit of two competitor models.

The method agrees with the minimalist identification philosophy underlying VARs. In fact, one can use some of the least controversial qualitative implications of a model to identify structural shocks in the data. Once shocks in data and the model are forced to have qualitatively similar features, the dynamic discrepancy between the two in the dimensions of interest can be easily examined. We summarize the main features of the approach in the next algorithm.

Algorithm 4.7

1) *Find qualitative, robust implications of a class of models.*

2) *Use (a subset of) these implications to identify shocks in the actual data. Stop validation if data does not conform to the qualitative robust restrictions of the model.*

3) *If theoretical restrictions have a data counterpart,* qualitatively *evaluate the model (use e.g. sign and shape of responses to shocks, the pattern of peak responses, etc.)*

4) *Validate* qualitatively *across models if more than one candidate is available.*

5) *If results in 3) and 4) are satisfactory, and policy analyses need to be performed, compare model and data* quantitatively*.*

6) *Repeat 2)-5) using other robust implications of the model(s), if needed.*

7) *If mismatch between theory and data is relevant, alter the model so as to maintain restrictions in 1) satisfied and repeat 3) and/or 5) to evaluate improvements. Otherwise, proceed to policy analyses.*

Few comments on algorithm 4.7 are in order. In 1) we require theoretical restrictions to be robust, that is independent of parametrization and/or of the functional forms of primitives. The idea is avoid restrictions which emerge only in special cases of the theory. In the second step we force certain shocks in the data and in the model to be qualitatively similar. In steps 2) to 7) evaluation is conducted at different levels: first, we examine whether the restrictions are satisfied in the data; second, we evaluate qualitative dynamic features of the model; finally, quantitative properties are considered. Qualitative evaluation should be considered a prerequisite to a quantitative one: many models can be discarded using the former alone. Also, to make the evaluation meaningful economic measures of discrepancy, as opposed to statistical ones, should be used.

The algorithm is simple, easily reproducible, and computationally affordable, particularly in comparison to ML or the Bayesian methods we discuss in Chapter 11; it can be used when models are very simplified descriptions of the actual data; and can be employed to evaluate one or more dimensions of the model. In this sense, it provides a flexible, limited information criteria which can be made more or less demanding, depending on the desires of the investigator. We illustrate the use of algorithm 4.7 in an example.

Example 4.37 *We take a working capital (WK) and a sticky price (SP) model, with the idea of studying the welfare costs of employing different monetary rules. We concentrate on the first step of the exercise, i.e. in examing which model is more appropriate to answer the policy question.*

Canova (2002) shows that these two models produce a number of robust sign restrictions in response to technology and monetary policy shocks. For example, in response to a policy disturbance the WK economy generates negative comovements of inflation and output, of inflation and real balances, and of inflation and the slope of the term structure and positive comovements of output and real balances. In the SP economy, the correlation between inflation and output is positive contemporaneously and for lags of output and negative for leads of output. The one between inflation and real balances is negative everywhere, the one of output and real balances is positive for lags of real balances and negative contemporaneously and for leads of real balances. Finally, the correlation of the slope of the term structure with inflation is negative everywhere. One could use some or all of these restrictions to characterize monetary shocks in the two models. Here we select restrictions on the contemporaneous cross correlation of output, inflation and the slope of the term structure for

the WK model and on the cross correlation of output, inflation and real balances in the SP model and impose them in a VAR composed of output, inflation, real balances, the slope of the term structure and labor productivity using US, UK and EURO data from 1980:1 to 1998:4.

We find that WK sign restrictions fail to recover monetary shocks in the UK, while SP sign restrictions do not produce monetary shocks in the Euro land. That is to say, out of 10000 draws for ω and $\mathcal{H}_{i,i'}(\omega)$ we are able to find less than 0.1% of teh cases where the restrictions are satisfied. Since no combination of reduced form residuals produces cross correlations for output, inflation and the slope (or real balances) with the required sign, both models are at odds with the dynamic comovements in response to monetary shocks in at least one data set. One may stop here and try to respecify the models, or proceed with the data sets where restrictions hold and evaluation can continue examining e.g. the dynamic responses of the two other VAR variables to identified monetary shocks.

There are at least two reasons for why a comparison based on real balances (or the slope) and labor productivity may be informative of the quality of the model's approximation to the data. First, we would like to know if identified monetary shocks produce liquidity effects, a feature present in both models and a simple "test" often used to decide whether a particular identification schemes is meaningful or not (see e.g. Leeper and Gordon (1994)). Second, it is common to use the dynamics of labor productivity to discriminate between flexible price real business cycle and sticky price demand driven explanations of economic fluctuations (see Gali (1999)). Since the dynamics of labor productivity in response to contractionary monetary shocks are similar in the two models (since employment declines more than output, labor productivity increases), it is interesting to check if the identified data qualitative conforms to these predictions.

Figure 4.6 plots the responses of these two variables for each data set (straight lines) together with the responses obtained in the two models (dotted lines), scaled so that the variance of the monetary policy innovation is the same. Two conclusions can be drawn. First, the WK identification scheme cannot account for the sign and the shape of the responses of labor productivity in US and Euro area and generates monetary disturbances in the Euro area which lack liquidity effects. Second, with the SP identification scheme monetary shocks generate instantaneous responses of the slope of the term structure which have the wrong sign in the US and lack persistence with UK data.

Given that the two theories produce dynamics which are qualitatively at odds with the data, it is not surprising to find that quantitative predictions are also unsatisfactory. For example, the percentage of output variance accounted for by monetary shocks in US at the 24 step horizon is between 11 and 43% with the WK scheme and 3 and 34% with the SP scheme. In comparison, and regardless of the parametrization used, monetary disturbances account for 1% of output variance in both models. Hence, both models lack internal propagation.

Given the mismatch of the models and the data one should probably go back to the drawing board before answering any policy question. Canova (2001) shows that adding capacity utilization and/or labor hoarding to the models is not enough to enhance at least the qualitative match. Whether other frictions will change this outcome is an open question.

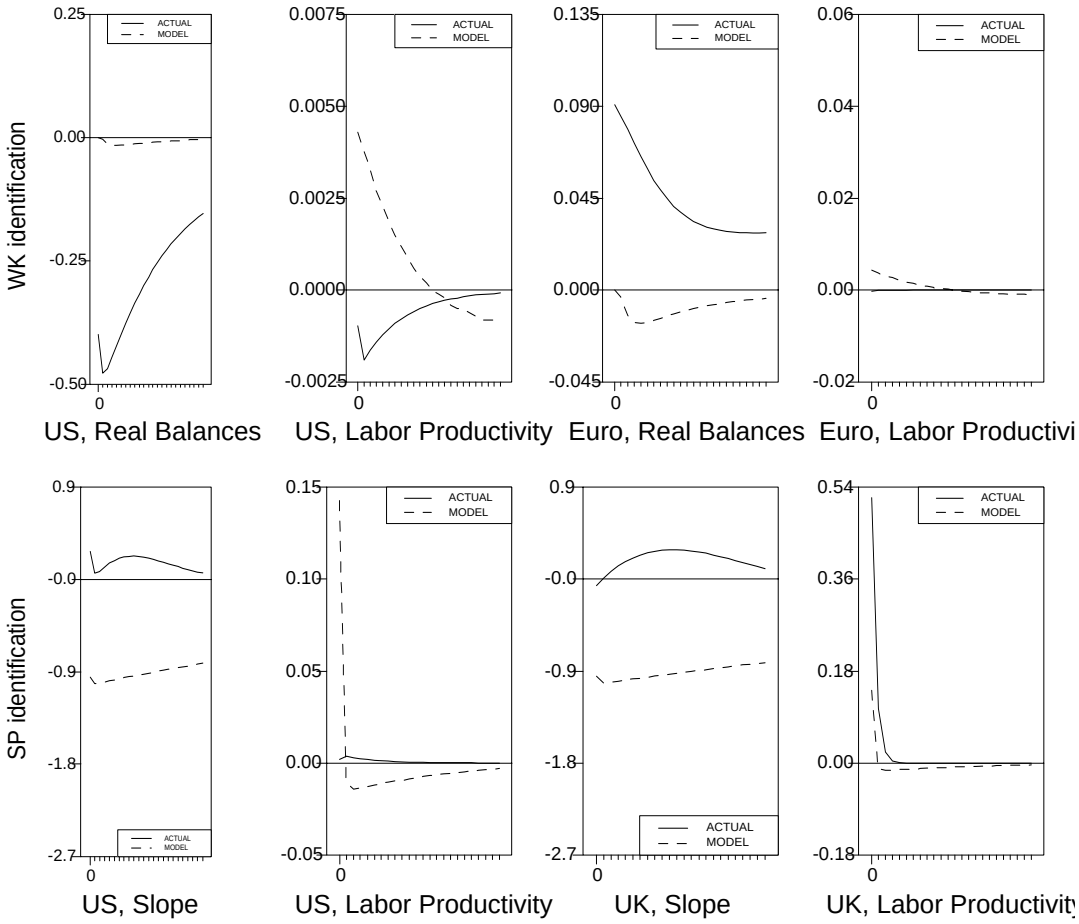


Figure 4.6: Responses to Monetary Shocks

Until a better match is found, it is probably unhealthy to try to answer any policy question with any of the two models.

Exercise 4.52 *(Dedola and Neri) Take a standard RBC model with habit persistence in consumption and highly persistent but stationary technology shocks. Examine whether robust sign restrictions for the correlation of output, hours and labor productivity exist when the extent of habit (γ), the power of utility parameter (φ), the share of hours in production (η), the depreciation rate (δ), and the persistence of technology shocks (ρ_ζ) are varied within reasonable ranges. Using a VAR with labor productivity, real wages, hours, investment, consumption and output examine whether the model fits the data, when robust sign restrictions are used to identify technology shocks in the data.*

Exercise 4.53 *(Pappa) In a sticky price model with monopolistic competitive firms anything that moves aggregate demand (e.g. goverment shocks) induces a shift in the labor demand curve and therefore induces positive comovements of hours and real wages. In a flexible price RBC model, on the other hand, goverment expenditure shocks shift both the aggregate supply and the aggregate demand curve. For many parametrizations movements in the former are larger than movements in the latter and therefore negative comovements of hours and real wages are generated. Using a VAR with labor productivity, hours, real wages, investment, consumption and output, verify whether a RBC style model fits the data better than a sticky price, monopolistic competitive model.*

Chapter 5: GMM and Simulation Estimators

A class of statistical and economic models feature orthogonality conditions of the form:

$$E[g(y_t, \theta) - \varrho] \equiv g_\infty(\theta) = 0 \tag{5.1}$$

where y_t is a $m \times 1$ vector of data observed at t, θ is a $k \times 1$ vector of parameters, g is a $n \times 1$ vector of functions and ϱ is a constant vector. Typically, E is the conditional expectation operator, i.e., $E[\cdot] \equiv E[\cdot|\mathcal{F}_t]$ where $\mathcal{F}_t$ is the information set at t. Sometimes, it represents unconditional expectations i. e. $E[\cdot] \equiv E(E[\cdot|\mathcal{F}_t])$.

Orthogonality conditions like (5.1) can be obtained from the first-order conditions of an intertemporal optimization problem, in which case θ contains preferences and technology parameters and y_t are the endogenous and the exogenous variables of the model; they emerge from some steady state relationship or appear among the identifying restrictions in (time series) regression models.

Whenever the model generating (5.1) is thought to represent the true data generating process (DGP), one can estimate θ using a variety of techniques. For example, noting that $E[g(y_t, \theta)] = g(y_t, \theta) - e_t$ where e_t is an expectational error, one could estimate θ by non-linear least square (NLLS) . Alternatively, one could use maximum likelihood (ML) once (a) the distributional properties of y_t are specified, and (b) an explicit closed-form solution, expressing the endogenous variables as a function of parameters and exogenous variables, is found. Clearly, in nonlinear models (b) is hard to obtain. Moreover, if n is large, both NLLS and ML may be computationally burdensome. Finally, distributional assumptions may be hard to justify if y_t includes endogenous variables.

The techniques we present in this chapter are designed to estimate θ and to test the validity of (5.1) without requiring either distributional assumptions or an explicit solution for the endogenous variables. The methodology can be applied to both linear and nonlinear specifications; it can be used for univariate ($n = 1$) and multivariate setups and requires only mild regularity conditions to produce estimators with "good" properties. One restriction we impose, at least in the initial framework, is that y_t only contains observable variables. Later, we relax it and allow some of the components of y_t to be unobservable.

The approaches we discuss are of limited information type. That is, estimation and testing is limited to conditions like (5.1). Therefore, although the model may have additional

equations, its degree of approximation to the data is examined only through the subset of its implications represented by (5.1). In the context of DSGE models this is not so restrictive: as we will see, optimality conditions will in fact imply restrictions like (5.1).

5.1 Generalized Method of Moment and other standard estimators

Definition 5.1 *Let $g_\infty(\theta)$ be the population mean of $g(y_t,\theta)-\varrho$, let $g_T(\theta)=\frac{1}{T}\sum_{t=1}^{T}(g(y_t,\theta)-\varrho)$ be its sample mean and let W_T be a symmetric positive definite $n\times n$ matrix. Then a Generalized Method of Moments (GMM) estimator θ_T solves*

$$arg\min_{\theta}(g_T(\theta_T)-g_\infty(\theta))'W_T(g_T(\theta_T)-g_\infty(\theta)) \tag{5.2}$$

A GMM estimator makes the sample version of certain orthogonality conditions "close" to their population counterpart and the matrix W_T describes what "close" means. GMM is therefore similar to a number of other estimators. For example, minimum distance estimators (see e.g. Malinvaud (1980)) also solve a problem like (5.2) but $g_T(\theta)-g_\infty(\theta)$ is not necessarily the difference between sample and population orthogonality conditions. Extreme estimators , i.e. estimators maximizing some criterion function (see Amemiya (1985)) are also obtained from a problem similar to (5.2).

Definition 5.1 includes several subcases of interest. For example, in many setups $g_\infty(\theta)=0$ in which case a GMM estimator sets the sample version of certain orthogonality conditions to zero. In other problems, expectations are conditional on time t information. Hence, if $E(g_{1t}(\theta))=0$, where g_1 is a scalar function, for any $z_t\in\mathcal{F}_t$ also $g_\infty(\theta)=E(z_tg_{1t}(\theta))=0$ in which case the solution to (5.2) produces a Generalized Instrumental Variable (GIV) estimator. Note that, if z_t is a constant, conditional and unconditional expectations are the same, so that GMM and GIV coincide.

We next present examples of economic models which deliver orthogonality conditions like (5.1) as part of the first order conditions of the problem.

Example 5.1 *Suppose a social planner maximizes $E_0\sum_t\beta^t u(c_t,(1-N_t))$ by choices of $\{c_t,N_t,K_{t+1}\}_{t=0}^{\infty}$ subject to $c_t+K_{t+1}\leq f(K_t,N_t)-G_t+(1-\delta)K_t$ where N_t are hours worked, K_t is capital and G_t is a random government expenditure disturbance. The first order conditions of the problem imply an Euler equation of the form*

$$E_t[\beta\frac{U_{c,t+1}}{U_{c,t}}[f_K+(1-\delta)]-1]=0 \tag{5.3}$$

where $f_K=\frac{\partial f}{\partial K}$, $U_{c,t}=\frac{\partial u}{\partial c_t}$. (5.3) fits (5.1) for $g(y_t,\theta)=\beta\frac{U_{c,t+1}}{U_{c,t}}(f_K+(1-\delta))$ and $\varrho=1$.

Example 5.2 *In the model of exercise 1.18 of chapter 2, the wage setting equation of monopolistic competitive workers was given by*

$$E_t \sum_{j=0}^{\infty} \beta^j \zeta_w^j (\frac{\pi^j w_t}{(1+\varsigma_w)p_{t+j}} U_{c,t+j} + U_{n,t+j}) N_{t+j} = 0 \tag{5.4}$$

where β is the discount factor, $U_{c,t+j}$ ($U_{n,t+j}$) is the marginal utility of consumption (labor) at $t+j$, p_t is the price level, π the steady state inflation rate, w_t the wage rate, ς_w is a parameter in the labor aggregator $N_t = (\int N_t(i)^{1/(1+\varsigma_w)} di)^{1+\varsigma_w}$, $i \in [0,1]$ and $1-\zeta_w$ is the fraction of workers allowed to change the wage each t. Then (5.4) fits (5.1) for $g(y_t,\theta) = \sum_{j=0}^{\infty} \beta^j \zeta_w^j (\frac{\pi^j w_t}{(1+\varsigma_w)p_{t+j}} U_{c,t+j} + U_{n,t+j}) N_{t+j}$ and $\varrho = 0$

Exercise 5.1 *Suppose agents maximize $E_0 \sum_t \beta^t u(c_t - \gamma c_{t-1})$ choosing $\{c_t, sa_{t+1}\}_{t=0}^{\infty}$, subject to the constraint $c_t + sa_{t+1} \leq w_t + (1+r)sa_t$ where w_t is an exogenous labor income, sa_{t+1} are savings maturing at $t+1$ and $r_t = r$, for all t. Show the orthogonality conditions of this problem. When is consumption a martingale process?*

Exercise 5.2 *A class of asset pricing models generates conditions like*

$$E_{t-1} r_{it} = rp_{0,t-1} + \sum_{j=1}^{J} \alpha_{ij} rp_{j,t-1} \tag{5.5}$$

$i = 0, 1, \ldots, m$, where r_{it} is the rate of return on asset i from $t-1$ to t, rp_{jt-1} are market wide expected risk premiums (conditional expected excess returns) and α_{ij} is the conditional beta of asset i relative to the $j-th$ "risk factor". Here rp_{jt-1} are latent variables.
i) Let $\tilde{r}_{it} = r_{it} - r_{0t}$, where r_{0t} is the return on a arbitrarily chosen asset. Show that (5.5) implies $E_{t-1}(\tilde{r}_t) = rp_{t-1}\theta$ where θ is a $J \times m$ matrix with $\theta_{ij} = \alpha_{ij} - \alpha_{0j}$. Show that, for any partition $\tilde{r} = (\tilde{r}_1, \tilde{r}_2)$, $E_{t-1}\tilde{r}_{2t}$ must be proportional to $E_{t-1}\tilde{r}_{1t}$.
ii) Show how to use (5.5) to set up orthogonality conditions to estimate the proportionality factor between expected returns of any two groups of assets.

Conditions like (5.1) are also common in consumption- based CAPM models.

Example 5.3 *Suppose agents maximize the same utility function as in example 5.1 choosing $\{c_t, B_{t+1}, S_{t+1}\}_{t=0}^{\infty}$ subject to $c_t + B_{t+1} + p_t^s S_{t+1} \leq y_t + (1+r_t)B_t + (p_t^s + sd_t)S_t$ where B_t are one period bond holdings, S_t are stock holdings, sd_t the dividends paid at t, r_t^B is the return on bonds and p_t^s the price of stocks at t. Optimality implies:*

$$E_t[\beta \frac{U_{c,t+1}}{U_{c,t}} \frac{p_{t+1}^s + sd_{t+1}}{p_t^s} - 1] = 0 \tag{5.6}$$

$$E_t[\beta \frac{U_{c,t+1}}{U_{c,t}} (1 + r_{t+1}^B) - 1] = 0 \tag{5.7}$$

where the first condition holds for stocks and the second for bonds. (5.6) and (5.7) fits (5.1) setting $g_1(y_t,\theta) = \beta \frac{U_{c,t+1}}{U_{c,t}} \frac{p_{t+1}^s + sd_{t+1}}{p_t^s}$ and $g_2(y_t,\theta) = \beta(1 + r_{t+1}^B) \frac{U_{c,t+1}}{U_{c,t}}$ and $\varrho_1 = \varrho_2 = 1$.

In sum, rational expectations models displaying some intertemporal link will generate at least one equation where a conditional expectation of some function of the variables is set to zero. Therefore, structures like (5.1) are pervasive in modern macroeconomics.

Many econometric and time series estimators can also be derived from orthogonality conditions like (5.1). We consider a few examples next.

Example 5.4 *Let* $f(y_t, \theta)$ *be the density of* y_t *and* θ *is a* $k \times 1$ *vector. Let* $E(y_t^i(\theta)) = \int y_t^i f(y_t, \theta) dy_t$, *and* $\hat{y}_T^i = \frac{1}{T}\sum_{t=1}^T y_t^i$ *be, respectively, the* $i-th$ *population and sample moment of* y_t. *A method of moment estimator* θ_{MM} *solves* $E(y_t^i(\theta)) = \hat{y}_T^i, i = 1, \ldots, k$. *Hence* $g_i(y_t, \theta) = [y_t^i - E(y_t^i(\theta)]$ *and* $\varrho = 0$.

Note that the estimator of example 5.4 requires k moments but it does not specify which ones a researcher should use. Since different moments produce different estimators, a method of moments estimator is not necessarily efficient. As we will see, a GMM estimator eliminates this source of inefficiency.

Example 5.5 *Let* $y_t = x_t\theta + e_t$ *with* $E_t[x_t' e_t] = 0$, *where* y_t *is a scalar and* x_t *a* $1 \times k$ *vector. Premultiplying by* x_t', *and taking conditional expectations we have* $E_t(x_t' y_t) = E_t(x_t' x_t)\theta + E_t(x_t' e_t)$. *Let* $y = (y_1, \ldots, y_t)'$ *and* $x = (x_1, \ldots, x_t)'$. *Then* $\theta_{OLS} = (x'x)^{-1}x'y$, *is a GMM estimator for* $g(x_t, \theta) \equiv x_t' y_t - x_t' x_t \theta = x_t' e_t$.

Exercise 5.3 *Suppose in example 5.5 that* $E_t(x_t' e_t) \neq 0$. *Let* z_t *be a set of instruments, correlated with* x_t *and satisfying* $E_t[z_t' e_t] = 0$. *Show the orthogonality conditions in this case. Show the g function that* θ_{IV} *solves.*

Example 5.6 *Suppose* $y_t = f(x_t, \theta) + e_t$. *Assume* $E_t[f(x_t, \theta)' e_t] \neq 0$ *and that there exists a set of* z_t *correlated with* x_t *such that* $E_t[z_t' e_t] = 0$. *Then* $E_t[z_t' e_t] = E_t[z_t'(y_t - f(x_t, \theta))] = E_t[g(y_t, z_t, x_t, \theta)] = 0$. *Therefore,* θ_{NLIV} *is a GMM estimator for* $g(y_t, z_t, x_t, \theta) = z_t'(y_t - f(x_t, \theta))$.

Exercise 5.4 *Consider a NLLS estimator of the model of example 5.6 when* $E_t(f(x_t)' e_t) = 0$. *Show the orthogonality conditions that NLLS solves.*

When g is linear in θ, a solution to the minimization problem is easy to find. When g is nonlinear, an estimator is found with an iterative procedure. An algorithm to find the estimator of example 5.6 is the following:

Algorithm 5.1

1) *Choose a* θ^0, *compute* $e(\theta^0) = y_t - f(x_t, \theta^0)$.

2) *Find* θ_1 *solving* $\frac{1}{T}\sum_t g_t = \frac{1}{T}\sum_t z_t' e(\theta^1) = 0$.

3) *Iterate on 1)-2) until* $||\theta^l - \theta^{l-1}|| < \iota$, ι *small,* $l = 2, 3, \ldots$.

Perhaps surprisingly, a ML estimator is also solves orthogonality conditions. Let $\{y_t\}_{t=0}^T$ be a stochastic process with density $f(y_t, \theta)$. Let $\mathcal{L}_T(\theta) = \sum_{t=1}^T \log f(y_t, \theta)$ be the sample log likelihood function for sample size T. If $\mathcal{L}_T(\theta)$ is strictly concave and differentiable, $\frac{\partial \mathcal{L}_T(\theta)}{\partial \theta} = 0$ is sufficient for the maximum. Then if $g(y_t, \theta) = \frac{1}{T}\frac{\partial \mathcal{L}_T(\theta)}{\partial \theta} = \frac{1}{T}\sum_{t=1}^T \frac{\partial \log f(y_t,\theta)}{\partial \theta}$, θ_{ML} is a GMM estimator. If $\mathcal{L}(\theta)$ is not globally concave, θ_{ML} obtained solving orthogonality conditions may be different from θ_{ML} obtained evaluating the likelihood directly, since the former is not designed to choose among local maxima.

Exercise 5.5 *Consider a $m \times 1$ VAR(q) model $y_t = A(\ell)y_{t-1} + e_t$ where $e_t \sim (0, \Sigma_e)$ and $E_t(e_t e'_{t-\tau}) = 0 \ \ \forall \tau \neq 0$. Show that $E_t(y_t y'_t) = \mathsf{A}E_t(y_{t-1}y'_t)$ where A is the companion of the matrix $A(\ell)$. Show the form of the g function θ_T will solve, where $\theta = vec(\mathsf{A}_1)$ and A_1 are the first m rows of A.*

While the class of GMM estimators is dense, there are less popular estimators that do not fit this framework. An example is given below

Example 5.7 *(Robust estimators) Consider the model of example 5.5, but suppose you want to neglect outliers. One way to do this is to minimize the sum of squares of e_t in a prescribed set, i.e $\min(\sum_t e_t^2) \times \mathcal{I}_{[\underline{e},\bar{e}]}$ where $\mathcal{I}_\cdot$ is an indicator function for the set $[\underline{e}, \bar{e}]$. The resulting trimmed estimator is generated from a g function with jumpy derivatives and violates one of the conditions given in section 5.3.*

5.2 IV estimation in a linear model

To understand the intuition behind GMM estimation it is useful to start from the problem of estimating regression parameters using instrumental variables. The intuition gained in this case carries over to more complicated nonlinear setups. Let $y = x\theta_0 + e, \ e \sim (0, \sigma^2 I)$ where y is a $(T \times 1)$ vector, x is a $(T \times k)$ stochastic matrix of rank k, θ_0 is a $(k \times 1)$ vector and e is a $(T \times 1)$ vector of disturbances. Let z be a $(T \times n)$ matrix of instruments satisfying $E(e_t|z_t) = 0$ (or $E(z'_t e_t) = 0$) $\forall t$ and let z_t be the t-th row of z (x_t could be an element of z_t). Let $e'z = \frac{1}{T}\sum_t e_t(\theta)' z_t$, where $e(\theta) = y - x\theta$, $z'x = \frac{1}{T}\sum_t z'_t x_t$, $z'y = \frac{1}{T}\sum_t z'_t y_t$, let $W_T \xrightarrow{a.s.} W$ be a $n \times n$ positive definite, symmetric matrix and define

$$Q_T(\theta) = [e(\theta)'z]W_T[e(\theta)'z]' \tag{5.8}$$

Let $\theta_{IV} = argmin\ [Q_T(\theta)]$. Taking the first-order condition of (5.8) with respect to θ and using the definition of $e(\theta)$ we have $x'zW_Tz'y = x'zW_Tz'x\theta$. We consider two cases $n = k$ and $n > k$. We neglect the third one, $n < k$, since in this case θ is underidentified, i.e., there is insufficient information to estimate θ uniquely. This is because the $(k \times k)$ matrix $x'zW_Tz'x$ will have rank $n < k$.

When $n = k$, the number of instruments is the same as the number parameters and $x'z$ is a square matrix. Hence, $\theta_{IV} = (z'x)^{-1}z'y = \theta_0 + (z'x)^{-1}z'e$ as long as $(z'x)$ is nonsingular. To show that θ_{IV} is consistent note that since $E(z'e) = 0 \ \forall t$, by the strong

law of large numbers we have that $z'e(\theta_0) \xrightarrow{a.s.} 0$. Also, since $W_T > 0$, $Q_T(\theta) \geq 0$. Hence, if $\theta_{IV} = arg\ min\ [Q_T(\theta)]$; $z'e(\theta_0) \xrightarrow{a.s.} 0$ and $z'x$ is bounded; it must be that $\theta_{IV} \xrightarrow{a.s.} \theta_0$. Clearly the argument breaks down if for some z_t, $E[z_t'e(\theta_0)] \neq 0$, i.e. the instruments are invalid. Here θ_{IV} does not depend on W_T.

When $n > k$, θ is overidentified, i.e., there is more information (orthogonality conditions) than it is necessary to estimate θ. In this case the choice of W_T matters. In fact, $x'zW_T$ is a $(k \times n)$ matrix so that $x'zW_Tz'e(\theta_{IV}) = 0$ does not necessarily imply that $z'e(\theta_{IV}) = 0$ but only that k linear combinations of the n orthogonality conditions $z'e(\theta)$ are set to zero with weights given by $x'zW_T$. The solution for θ_{IV} is

$$\theta_{IV} = (x'zW_Tz'x)^{-1}x'zW_Tz'y \tag{5.9}$$

Exercise 5.6 *Give sufficient conditions to insure that $\theta_{IV} \xrightarrow{a.s.} \theta_0$ in (5.9).*

To characterize the asymptotic distribution of θ_{IV} when $n \geq k$, use the model and (5.9) to obtain

$$\sqrt{T}(\theta_{IV} - \theta_0) = (x'zW_Tz'x)^{-1}x'zW_T\sqrt{T}z'e \tag{5.10}$$

We make three assumptions: (i) $\lim_{T\to\infty}\ z'z = \Sigma_{zz}$, $|\Sigma_{zz}| \neq 0$; (ii) $\lim_{T\to\infty}\ x'z = \Sigma_{xz}$, $rank\ |\Sigma_{xz}| = k$; (iii) $\lim_{T\to\infty}\ \sqrt{T}z'e \xrightarrow{D} \mathsf{N}(0, \sigma^2\Sigma_{zz})$. The first condition requires that each instrument provides unique information; the second that at least k instruments correlate with the x's; the third that the scaled sample orthogonality conditions evaluated at θ_0 converge to a normal distribution.

Exercise 5.7 *Using the three above conditions, show that (5.10) implies*

$$\sqrt{T}(\theta_{IV} - \theta_0) \xrightarrow{D} \mathsf{N}(0, \Sigma_\theta) \tag{5.11}$$

where $\Sigma_\theta = (\Sigma_{xz}W\Sigma_{zx})^{-1}\Sigma_{xz}W\sigma^2\Sigma_{zz}W'\Sigma_{zx}((\Sigma_{xz}W\Sigma_{zx})^{-1})'$ and $\Sigma_{zx} = \Sigma_{xz}'$. Show that when $k = n$, the expression simplifies to $\Sigma_\theta = (\Sigma_{xz}^{-1}\sigma^2\Sigma_{zz}\Sigma_{zx}^{-1})$.

Exercise 5.8 *Suppose that $rank\ |\Sigma_{xz}| < k$. What does this tells you about your instruments? Would a IV approach work? What would the distribution in (5.11) be?*

To summarize, θ_{IV} minimizes a quadratic form in $z'e(\theta)$. The estimator is consistent because $E(z'e(\theta)) = 0$ and asymptotically normal because all quantities converge to non-stochastic matrices. These same principles underlie GMM estimation and the proofs of its asymptotic properties.

Since both θ_{IV} and Σ_θ depend on W, it is natural to ask which W will lead to the most efficient θ_{IV}, i.e the one which minimizes Σ_θ. Once such a W is found, W_T could be any sequence of matrices converging to W almost surely.

Exercise 5.9 *Show that a solution to $\min_W \Sigma_\theta(W)$ is $W^\dagger = \sigma^{-2}\Sigma_{zz}^{-1}$ and that $\Sigma_\theta^\dagger \equiv \Sigma_\theta(W^\dagger) = \Sigma_{zx}^{-1}\sigma^2\Sigma_{zz}\Sigma_{xz}^{-1}$ (Note that this is only sufficient for efficiency. Why?).*

The optimal weighting matrix is proportional to the asymptotic covariance matrix of the instruments. To gain intuition into this choice, note that different instruments contain different information about θ (because they have different variabilities). The optimal weighting matrix gives less weight to instruments which are very volatile. When $n > k$, one can choose which conditions to use. If the first k restrictions are employed the estimator (say, $\theta_{1(IV)}$) would be numerically different but it will have the same asymptotic properties of $\theta_{2(IV)}$ obtained using, say, the last k conditions. The optimal weighting matrix combines the information contained in all conditions and maximizes efficiency.

Exercise 5.10 *Consider the money demand function (derived from a cash-in-advance model)* $\frac{M_t}{p_t} = GDP_t\theta + e_t$ *where* θ *is the inverse of the constant velocity and* e_t *appears because* GDP_t *could be measured with error or because variables (for example, the nominal interest rate) are omitted. Since current GDP could be correlated with* e_t*, consider two sets of instruments* $z_t^1 = [GDP_{t-1}]$ *and* $z_t^2 = [GDP_{t-1}, GDP_{t-2}]$*. Show that* θ_{IV}^2 *obtained using* z_t^2 *is at least as efficient as* θ_{IV}^1 *obtained using* z_t^1*. Give some intuition for why it is the case and conditions under which the asymptotic covariance matrix of the two estimators is identical.*

When $W^\dagger$ is used, θ_{IV} becomes:

$$\theta_{IV}^\dagger = (\hat{x}'\hat{x})^{-1}\hat{x}'y \tag{5.12}$$

where $\hat{x} = z(z'z)^{-1}z'x$. Consistent estimators of σ^2 and Σ_θ can be constructed using $\hat{\sigma}^2 = \frac{(y-x\theta_{IV}^\dagger)'(y-x\theta_{IV}^\dagger)}{T}$ and the sample matrices $\frac{1}{T}\sum_t z_t'x_t$ and $\frac{1}{T}\sum_t z_t'z_t$. Note that we need an iterative approach to compute $W^\dagger$ since it depends on $\theta_{IV}^\dagger$ via $\hat{\sigma}^{-2}$and $\theta_{IV}^\dagger$ depends on $W^\dagger$ via $W_T^\dagger$. A standard way to proceed is to choose W_T suboptimally, e.g., $W_T = I$ or $W_T \propto \frac{1}{T}\sum_t z_t'z_t$ and obtain a θ_{IV}^1 which is consistent although inefficient. Given θ_{IV}^1 we can construct $W_T^\dagger$ and obtain θ_{IV}^2. Under regularity conditions θ_{IV}^2 will be equivalent to a fully iterative θ_{IV}.

Exercise 5.11 *Consider the linear model*

$$y = x\theta_1 + e \tag{5.13}$$
$$x = z\theta_2 + v \tag{5.14}$$

where y, x, e, v *are* $T \times 1$ *vectors,* z *is a* $T \times n$ *matrix,* θ_2 *is a* $n \times 1$ *vector and* θ_1 *a scalar. Suppose that* $E(v|z) = 0$*, let* $\theta_{1,ols} = (x'x)^{-1}(x'y)$ *and* $\theta_{1,IV} = (\hat{x}'\hat{x})^{-1}(\hat{x}y)$*.*
(i) Show $p\lim\theta_{1,ols} = \theta_1 + \frac{cov(x,e)}{var(x)}$*;* $p\lim\theta_{1,IV} = \theta_1 + \frac{cov(\hat{x},e)}{var(\hat{x})}$*. Argue that unless* $E(e|z) = 0$ *and* $E(x|z) \neq 0$*,* $p\lim\theta_{1,IV}$ *may not exist.*
(ii) Show that the inconsistency of $\theta_{1,IV}$ *relative to* $\theta_{1,OLS}$ *is* $RI = \frac{cov(\hat{x},e)}{cov(x,e)}/R_{xz}^2$ *where* R_{xz}^2 *is the regression* R^2 *in (5.14)*
(iii) Let $n = 1$*. Show that if* z *is weakly correlated with* x*,* $RI \to \infty$ *even if* $E(e|z) \approx 0$*.*
(iv)An approximate bias for θ_{IV} *is* $\frac{cov(e,v)}{var(v)}\frac{var(v)}{\theta_2'z'z\theta_2}(n-2)$*. Argue that the bias of* $\theta_{1,IV}$ *relative*

to the $\theta_{1,OLS}$ is inversely proportional to the F-statistics on the instruments in (5.14) so that weakly cotrrelated instruments will send the relative bias to infinity. Propose a statistic to test for the IV bias and suggest a sequential procedure to choose instruments.

Exercise 5.12 *Show that a 2SLS estimator of the form $\theta_{2SLS} = [(x'z)(z'z)^{-1}(z'x)]^{-1}$ $[(x'z)(z'z)^{-1}(z'y)]$ corresponds to the optimal estimator derived in (5.12). Display the orthogonality conditions and the g function in this case.*

Extensions of these concepts to an $m \times 1$ vector of equations are straightforward and left as exercises for the reader.

Exercise 5.13 *Show that $\theta_{3SLS} = [(\mathsf{X}'(I \otimes \mathsf{Z}')')(\Sigma_e^{-1} \otimes (\mathsf{Z}'\mathsf{Z})^{-1})((I \otimes \mathsf{Z}')\mathsf{X})]^{-1}$ $[(\mathsf{X}'(I \otimes \mathsf{Z}')')((\Sigma_e^{-1} \otimes \mathsf{Z}'\mathsf{Z})^{-1}((I \otimes \mathsf{Z}')y)]$ is the optimal estimator obtained from the $mn \times 1$ orthogonality conditions $E((I \otimes \mathsf{Z})'e) = 0$ where $E(ee'|z) = \Sigma_e$ is a $m \times m$ matrix and the same Z matrix is used in all equations. Show that $var(\theta_{3SLS}) = [(\mathsf{X}'(I \otimes \mathsf{Z}')')(\Sigma_e^{-1} \otimes$ $(\mathsf{Z}'\mathsf{Z})^{-1})((I \otimes \mathsf{Z}')\mathsf{X})]^{-1}$. Provide conditions for consistency and asymptotic normality of θ_{3SLS}.*

Example 5.8 *As seen in chapter 4, a VAR is a particular simultaneous equation system with the same regressors in every equation. In this case the orthogonality conditions of exercise 5.13 are of the form $E(z_{ji}e_{ji'}) = 0 \quad \forall i, i' = 1, \ldots, m$, since regressors are orthogonal across equations. Hence z_i becomes irrelevant and the optimal estimator is $\theta_{SUR} = (\mathsf{X}'(\Sigma_e^{-1} \otimes I)\mathsf{X})^{-1}(\mathsf{X}'(\Sigma_e^{-1} \otimes I)y)$.*

So far we have assumed that e_t are conditional homoskedastic. In some applications it is more reasonable to assume that $E_t[z_t'e_te_t'z_t]$ can not be factored into the product of σ^2 and of Σ_{zz}. Most of the arguments made go through if $E_t[z_t'e_te_t'z_t] = \Sigma_{ez}$. However, to prove consistency and asymptotic normality we need to strengthen the assumptions to include the condition that $E(z'xx'z)$ exists and it is finite (see Hayashi (2002, p. 212)).

Exercise 5.14 *Suppose $\lim_{T\to\infty} \sqrt{T}z'e \xrightarrow{D} \mathsf{N}(0, \Sigma_{ez})$. Derive the distribution of the optimal IV estimator. Provide an estimate for $W_T^\dagger$ and show the form of $\Sigma_\theta^\dagger$.*

Example 5.9 *A widely used statistical model with conditional heteroskedasticity is a regression model with GARCH errors. Here $y = x\theta_0 + e$ and $var(e_t) \equiv \sigma_t^2 = b_1\sigma_{t-1}^2 + e_t^2 + b_2e_{t-1}^2$. Since σ_t^2 depends on the level of the regressors and since the instruments are correlated by construction with the level of the regressors, $E(z_t'e_t)(z_t'e_t)' = \Sigma_t$ and Σ_t is serially correlated.*

Exercise 5.15 *Suppose that in the log-linearized version of money demand equation used in exercise 5.10 we had erroneously used GDP deflator (GDPD) in place of CPI for measuring prices. In that case $e_t = log(CPI_t) - log(GDPD_t)$. Suppose a researcher estimates $\log M_t = \theta_1 \log GDPD_t + \theta_2 \log GDP_t + e_t$ and guessing that some misspecification occurs, estimates $\theta = (\theta_1, \theta_2)$ using one lag of $GDPD_t$ and of GDP_t as instruments. Show the conditions under which the orthogonality conditions will display conditional heteroskedasticity. Are there conditions that make the orthogonality conditions serially uncorrelated?*

In some applications, the condition that $E_t(g(y_t,\theta))=0$ is hard to maintain. For example, e_t may be serially correlated in which case $E(g_t g_{t-j}) \neq 0,\ j=1,\ldots$. The asymptotic distribution and the consistency proof are unchanged by this alteration. However, as in regression models with serially correlated errors, the asymptotic covariance matrix needs to be modified. We defer the presentation of such covariance matrix to a later section.

Example 5.10 *Consider the problem of a representative agent who chooses how much to consume and save and at what maturities to lock her savings in. Assume that there are only one and τ periods government bonds, issued in fixed supply every period, paying $(1+r_{jt})$ at time $t, j=1,\tau$. The Euler equations are*

$$E_t[\beta\frac{U_{c,t+1}}{U_{c,t}}-\frac{1}{1+r_{1t}}] = 0 \tag{5.15}$$

$$E_t[\beta^{\tau}\frac{U_{c,t+\tau}}{U_{c,t}}-\frac{1}{1+r_{\tau t}}] = 0 \tag{5.16}$$

These two conditions imply that the expected (at t) forward rate for $\tau-1$ periods must satisfy the no arbitrage condition $E_t[\beta^{\tau-1}\frac{U_{c,t+\tau}}{U_{c,t+1}}-\frac{1+r_{1t}}{1+r_{\tau t}}]=0$. Log linearizing around the steady state, assuming a separable log utility and letting $y_{t+\tau}=-\hat{c}_{t+\tau}+\hat{c}_{t+1},\ x_t=-\frac{r_\tau}{1+r_\tau}\hat{r}_{\tau t}+\frac{r_1}{1+r_1}\hat{r}_{1t}$ where $\hat{\cdot}$ represents percentage deviations from the state, $y_{t+\tau}=\theta x_t+e_{t+\tau}$ and $\theta_0=1$. Note that $E_t[e_{t+\tau}]=0$ but that unless the sampling interval of the data exactly equals the forward rate interval, $e_{t+\tau}$ will be serially correlated. For example, if data is monthly and τ is 12, there will be moving average terms of order 11.

Exercise 5.16 *Suppose analysts are asked each quarter to produce output forecasts τ periods ahead and suppose an investigator is asked to evaluate whether the forecasts are rational or not. Let $y_{t+\tau}$ be realized output at $t+\tau$ and $y_t(\tau)$ the forecast at t of $y_{t+\tau}$.*
(i) Show that rationality implies orthogonality conditions with moving average terms of order up to τ.
(ii) Explain why GLS estimates obtained from $y_{t+\tau}=\theta_1+\theta_2 y_t(\tau)+e_{t+\tau}$ are inconsistent.
(iii) Show that the asymptotic covariance matrix for $\theta=(\theta_1,\theta_2)$ is $(\frac{1}{T}\sum_t z_t'x_t)^{-1}$ $(\frac{1}{T}\sum_t z_t'\Sigma z_t)(\frac{1}{T}\sum_t x_t'z_t)$ where z_t are instruments, $\Sigma=\{\sigma_{ij}\}$ and $\sigma_{ij}=\sigma^2 ACF(|i-j|)$ for $|i-j|<\tau$ and zero otherwise and $ACF(|i-j|)$ is the $(i-j)$-th element of the autocovariance function of $e_{t+\tau}$

Exercise 5.17 *In the context of example 5.10 consider the question of pricing a "crop insurance". Let $p_{t,\tau}(c_t,\theta)$ be the price at t in terms of consumption goods of a claim to consumption at $t+\tau$ if the crop c_t falls below θ. It is easy to verify that $p_{t,\tau}(c_t,\theta)=\beta^t\int_0^\theta \frac{U_{c,t+\tau}}{U_{c,t}}P_{t,t+\tau}dc_{t+\tau}$ where $P_{t,t+\tau}$ is the transition probability from c_t to $c_{t+\tau}$. Show the pricing formula for a crop insurance of maturity up to 2. Conclude that the Euler equation for pricing crop insurance up to maturity τ will have errors with up to $\tau-1$ MA components.*

Exercise 5.18 *Suppose the orthogonality conditions are serially correlated, but an investigator neglects to take this into account. Is the optimal IV still consistent? Is it efficient?*

Before testing hypotheses on θ, one may want to check whether the orthogonality conditions are correctly specified. For a simple regression setup many tests for adequacy exist, e.g., tests for serial correlation, for heteroskedasticity, etc. In general setups, including those implied by DSGE models , the assumption that e_t are serially uncorrelated or homoskedastic can not be made. Hence, we need a procedure to check model adequacy without focusing on these "statistical" features.

When $n = k$, $z'e(\theta) = 0$ by construction, so no test is possible. When $n > k$, only k linear combinations of $z'e(\theta)$ are set to zero so that $z'e(\theta_{IV})$ may differ from zero. However, if the population conditions are true, one should expect $z'e(\theta_{IV}) \approx 0$. Hence a specification test under the null that $E[z'e(\theta_0)] = 0$, is given by $T \times \{[e(\theta_{IV}^{\dagger})'z][\tilde{\sigma}^2\Sigma_{zz}]^{-1}[z'e(\theta_{IV}^{\dagger})]\} \xrightarrow{D} \chi^2(n-k)$. (This is typically called $J-$test). Since k conditions are used to obtain $\theta_{IV}^{\dagger}$, $n-k$ conditions are left free for testing and the number of degrees of freedom of the test equals the number of overidentifying restrictions.

Consistent estimators can be used in place of the true σ^2 and Σ_{zz} without changing the distribution of the tests.

Tests of hypotheses on θ can be conducted in standard ways. We defer the discussion of general testing to a later section.

Example 5.11 *In exercise 3.12 of chapter 2 we have asked the reader to verify that a log-linearized version of a New-Keynesian Phillips curve is*

$$\pi_t = \beta E_t \pi_{t+1} + \kappa \frac{(1-\zeta_p)(1-\zeta_p\beta)}{\zeta_p} mc_t \tag{5.17}$$

where $mc_t = \frac{N_t w_t}{GDP_t}$ *are real marginal costs,* ς_p *is the probability of not changing the prices,* κ *is a function of the risk aversion parameter and the labor supply elasticity and* π_t *is the inflation rate. Clearly (5.17) is an orthogonality condition. In earlier works, the marginal cost was proxied by output gap and the parameters of* $\pi_{t+1} = \alpha_1\pi_t - \alpha_2 gap_t + e_{t+1}$ *where* e_t *is an expectational error,* gap_t *is the difference between output and its potential output, were estimated using* $z_t = (\pi_t, gap_t)$ *as instruments. Fixing* $\beta = 1$*, Gali and Gertler (1999) found* α_2 *to be negative and significant, contrary to the theory. However, when marginal costs were proxied by the labor share, estimates of* α_2 *were positive and significant.*

We estimate (5.17) using CPI inflation data for US, UK and Germany for the sample 1980:1-2000:4 proxing marginal costs with the output gap (computed using a HP filter) or with unit labor costs. The first two columns of table 5.1 show that Gali and Gertler's conclusions roughly hold, even though the coefficient of the labor share in the US is only marginally significant (t-tests are in parenthesis). Letting $\kappa = 1$*, columns 3 and 4 report IV estimates of* ζ_p *and* β *obtained using CPI inflation for the three countries. Instruments include a constant, lags of* π *and lags of the output gap (or of the labor share). In each case we report estimates which are most favorable to the theory. In many cases estimates obtained from just or overidentified specifications, where the optimal weighting matrix is used, are similar. Three main results stand out. First, when output gap is used, estimates of* ζ_p *are reasonable for US and UK (they imply, on average, slightly less than 3 quarters*

	Reduced Form		*Structural*		
Country	α_1	α_2	β	ζ_p	*J-test p-value*
US-Gap	*0.993(16.83)*	*-0.04 (-0.31)*	*0.907 (10.35)*	*0.700 (5.08)*	$\chi^2(5)$=*0.15*
US-LS	*0.867(10.85)*	*0.001(1.75)*	*0.932 (7.74)*	*0.991 (150.2)*	$\chi^2(9)$=*0.54*
UK-Gap	*0.667(7.18)*	*0.528(1.30)*	*0.924 (4.96)*	*0.684 (1.37)*	*NA*
UK-LS	*0.412(3.81)*	*0.004(4.05)*	*0.853 (4.07)*	*0.994 (166.1)*	$\chi^2(1)$=*0.25*
GE-Gap	*0.765(10.02)*	*-0.01 (-0.22)*	*0.972 (7.48)*	*1.014 (0.03)*	*NA*
GE-LS	*0.491(3.34)*	*0.03 (1.85)*	*0.919 (4.45)*	*0.958 (7.18)*	$\chi^2(1)$=*0.83*

Table 5.1: Estimates of New Keynesian Phillips curve

between price changes) but not for Germany. Second, when labor share is used estimates of ζ_p are close to 1, except for Germany, while estimates of β are smaller indicating that the two parameters may not be separately identifiable. Finally, despite the poor structural estimates, the model's orthogonality conditions are not rejected. One reason for why this happens could be that the instruments are poor.

5.3 GMM Estimation: An overview

The machinery described in section 2 can be applied with slight variations to setups where the orthogonality conditions are nonlinear in θ.

Assume that g_t satisfies $E(E_t[g(y_t,\theta_0)]) = E[g(y_t,\theta_0)]$. Let $g_T(\theta) = \frac{1}{T}\sum_{t=1}^{T} g(y_t,\theta)$ and $h_T(\theta) = g_T(\theta) - g_\infty(\theta)$. Then the θ_T which minimizes $Q_T(\theta) = h_T(\theta)'W_T h_T(\theta)$ solves

$$H_T(\theta_T)'W_T h_T(\theta_T) = 0 \tag{5.18}$$

where $H_T(\theta_T)$ is a $n \times k$ matrix of rank k, $[H_T(\theta_T)]_{ij} = \partial h_{T_i}(\theta_T)/\partial\theta_j$ where $h_{T_i}(\theta_T)$ is the i^{th} element of $h_T(\theta_T)$. When $n = k$, $H_T(\theta_T)$ and W_T are nonsingular and θ_T solves $h_T(\theta_T) = 0$. When $k < n$, θ_T depends on W_T. To derive the asymptotic distribution of θ_T, we need a closed form solution for θ_T which, in general, is unavailable. Using the mean value theorem we can write (5.18):

$$(\theta_T - \theta_0) = -[H_T(\bar\theta)'W_T H_T(\overline{\theta})]^{-1} H_T(\bar\theta)'W_T h_T(\theta_0) \tag{5.19}$$

where $\bar\theta \in [\theta_0, \theta_T]$. To show that θ_T is consistent, the expression on the right hand side of (5.19) must go to zero almost surely or in probability. To prove asymptotic normality we need to make assumptions on h_T so that $\lim_{T\to\infty}\sqrt{T}h_T(\theta_0) \xrightarrow{D} \mathsf{N}(0,\Sigma_h)$ and make sure that the other quantities in the right hand side of (5.19) have finite non-random limiting behavior. Then the asymptotic covariance matrix of θ_T is a multiple of Σ.

As in section 2, the optimal W_T minimizes the asymptotic covariance matrix of θ_T. Also here the computation of $W_T^\dagger$ is complicated by the fact that θ_T depends on $W_T^\dagger$ and $W_T^\dagger$ can be computed only if θ_T is known. With a large T, a two-step GMM is as efficient as

fully iterative GMM if the first step estimate converges to the true parameter at the rate $\sqrt{T}$ ($\sqrt{T}$-consistency). In small samples, iterative estimators may be more accurate.

To perform tests on θ_T we need to have a consistent estimate of the asymptotic covariance matrix. While $W_T \xrightarrow{a.s} W$ by construction, and consistent estimators of H_T can easily be obtained, some care is needed to get a consistent estimator of Σ_h which is positive semi-definite.

As in linear models, when θ is overidentified $h_T(\theta_T) \neq 0$, but if θ_T is "correct" $h_T(\theta_T) \approx 0$. Hence, under the null that $E[h(y_t, \theta_0))] = 0$, $T \times h_T(\theta_T)' \Sigma_h^{-1} h_T(\theta_T) \xrightarrow{D} \chi^2(n-k)$. General hypothesis on θ_T can be tested using Wald, Lagrange or distance tests. In the next subsections we make these arguments more precise.

5.3.1 Asymptotics of GMM estimators

The discussion here is sketchy and brief. For a more thorough presentation the reader should refer to Gallant (1987) or Newey and Mc Fadden (1994). The conditions are general and need to be specialized when applied to orthogonality conditions derived from stationary DSGE models. Throughout this subsection we assume that there exists a θ_0 such that $h_T(\theta_0) \xrightarrow{a.s.} 0$ as $T \to \infty$ and that $\sqrt{T} h_T(\theta_0) \xrightarrow{D} \mathsf{N}(0, \Sigma_h)$. Intuitively the first condition implies strong ergodicity of h_T; the second asymptotic normality of the difference between sample and population g functions.

To prove consistency we need three assumptions. First, that $\theta \in \Theta$, a closed and bounded set. Second, that $h_T(\theta)$ is continuous and converges uniformly to $h(\theta)$ on Θ, almost surely. Third, that θ_0 is the unique solution to $h_T(\theta) = 0$. Then the proof requires $Q_T(\theta)$ to converge uniformly on Θ to some $Q(\theta)$ which has a unique minimum at θ_0. The above conditions imply that this is the case. In fact, by the second assumption $Q_T(\theta)$ converges uniformly over Θ to $Q(\theta) = h'(\theta) W h(\theta)$. Since we have assumed that $\lim_{T\to\infty} h_T(\theta_0) \xrightarrow{a.s.} 0$ we have that $h(\theta_0) = 0$. Then, if θ_0 is the unique solution of $h(\theta) = 0$ over Θ, $Q(\theta)$ has a unique maximum and $\lim_{T\to\infty} \theta_T \xrightarrow{a.s.} \theta_0$.

Uniform convergence of $h_T(\theta)$ is hard to verify in practice. Therefore, one typically assumes that $E[sup_\theta ||h(y_t, \theta)||] < \infty$ - a condition easier to check. This together with continuity of g in θ, measurability of g in y_t (i.e. g must be continuous in y_t for each θ) and the uniform law of large numbers can be used to show that $h_T(\theta) \to h(\theta)$ uniformly. Compactness of Θ is also hard to obtain in practice (it requires knowing the upper and the lower bound of the parameters space). The alternative assumption typically made - that the objective function is concave together with pointwise convergence of $h_T(\theta)$ to $h(\theta)$ (see Newey and McFadden (1994), p. 2133) is unappealing since in general $h_T(\theta)$ may not be concave. Another alternative (see Gallant (1987)) is to impose restrictions on the tails of the distribution of h_T and to show that, for large T, they imply that θ is in a closed and bounded set for all $t \geq T$.

To specialize this theorem to the setup generated by DSGE models, we need to insure $h_T(\theta_0) \xrightarrow{a.s.} 0$ and that $\sqrt{T} h_T(\theta_0) \xrightarrow{D} \mathsf{N}(0, \Sigma_h)$. If y_t is stationary and ergodic and if g is continuous, then g_t is also stationary and ergodic. Furthermore, the sum of stationary and

ergodic processes is also stationary and ergodic so that $h_T(\theta_0) \xrightarrow{a.s.} 0$. If y_t is stationary and g_t is a martingale difference, the martingale central limit theorem insures that $\sqrt{T}g_T(\theta_0) \xrightarrow{D} \mathsf{N}(0, var[g(y_t, \theta_0)])$ and, if $g_\infty = 0$, $\sqrt{T}h_T(\theta_0) \xrightarrow{D} \mathsf{N}(0, \Sigma_h)$. Recall that the martingale difference assumption requires only first moment independence. Hence, it is a weaker condition than simple independence.

Exercise 5.19 *Let $y_t = x_t\theta + e_t$ where (y_t, x_t) are jointly stationary and ergodic sequences, $E(x_t'e_t) = 0$, $E(x_t'x_t) = \Sigma_{xx} < \infty$, $|\Sigma_{xx}| \neq 0$. Show $\theta_{OLS} \xrightarrow{a.s.} \theta_0$. How does the proof differ from the one where observations are iid?*

In certain applications the assumption of identically (homogeneously) distributed time series is implausible (can you give an example when this is the case?). Then, it is typical to substitute the stationarity-ergodicity assumptions for y_t with some mixing requirement.

Exercise 5.20 *State conditions and prove the consistency of $\theta_{2SLS} = [x'z(z'z)^{-1}z'x]^{-1}$ $(x'z)(z'z)^{-1}z'y)$ in the model $y_t = x_t\theta + e_t$, where x_t is a $(1 \times k)$ vector; z_t is a $(1 \times n)$ $k < n$ and when y_t, x_t, z_t satisfy α-mixing conditions.*

To show asymptotic normality we need two extra assumptions: that θ_0 is in the interior of Θ and that $H_T(\theta)$ is continuous and converges uniformly to $H(\theta) = \frac{\partial h(\theta)}{\partial \theta'}$ (in addition to the fact that $\theta_T \xrightarrow{a.s.} \theta_0$). The first assumption is needed to make the Taylor expansion presented below well behaved. The second insures that the partial derivatives of the h_T function carry proper information for θ. For the typical row of $h_T(\theta)$ we have

$$\sqrt{T}h_{i,T}(\theta_T) = \sqrt{T}h_{i,T}(\theta_0) + \frac{\partial h_{iT}(\bar{\theta})}{\partial \theta'}\sqrt{T}(\theta_T - \theta_0) \tag{5.20}$$

where $\bar{\theta} \in [\theta_0, \theta_T]$. Because $\bar{\theta}$ is in the line segment joining θ_0 and θ_T, and because $\theta_T \xrightarrow{a.s.} \theta_0$ also $\bar{\theta} \xrightarrow{a.s.} \theta_0$. Moreover, given the assumptions, the typical row of $H_T(\theta)$ converges uniformly over Θ to $H(\theta_0)$, i.e. $\frac{\partial h_{i,T}(\bar{\theta})}{\partial \theta'} \xrightarrow{a.s.} \frac{\partial h_i(\theta_0)}{\partial \theta'}$. Applying the argument to each row we have that $\sqrt{T}h_T(\theta_T) = \sqrt{T}h_T(\theta_0) + H(\theta_0)\sqrt{T}(\theta_T - \theta_0)$. Substituting in (5.18) we have

$$0 = H_T(\theta_T)'W_T\sqrt{T}h_T(\theta_0) + H_T(\theta_T)'W_TH(\theta_0)\sqrt{T}(\theta_T - \theta_0) \tag{5.21}$$

Because $H_T(\theta_T)'W_T \xrightarrow{a.s.} H(\theta_0)'W$ and $\sqrt{T}h_T(\theta_0) \xrightarrow{D} \mathsf{N}(0, \Sigma_h)$ we have that $H_T(\theta_T)'W_T$ $\sqrt{T}h_T(\theta_0) \xrightarrow{D} \mathsf{N}(0, H(\theta)'W\Sigma_hW'H(\theta))$. Furthermore $H_T(\theta_T)'W_TH(\theta_0) \xrightarrow{a.s.} H(\theta_0)'WH(\theta_0)$. Therefore $\sqrt{T}(\theta_T - \theta_0) \xrightarrow{D} N(0, \Sigma_\theta)$ where $\Sigma_\theta = (H(\theta_0)'WH(\theta_0))^{-1}$ $H(\theta_0)'W\Sigma_hW'H(\theta_0)$ $(H(\theta_0)'WH(\theta_0)^{-1})'$ which simplifies to $\Sigma_\theta = (H(\theta_0)'\Sigma_h^{-1}H(\theta_0))^{-1}$ when $W = \Sigma_h^{-1}$.

To intuitively understand the proof of asymptotic normality consider figure 5.1

Under certain assumptions we can translate knowledge of $h_T(\theta)$ in information about $\theta_T - \theta_0$. For this to happen we need to impose regularity conditions on h_T so that the mapping is well defined. In particular, we need that (i) h_T is differentiable for all y_t and that (ii) H_T is such that $H_T(\theta_T)$ and $H_T(\theta_0)$ do not differ too much. While these conditions appear to be generally satisfied, it is easy to build examples where they are not.

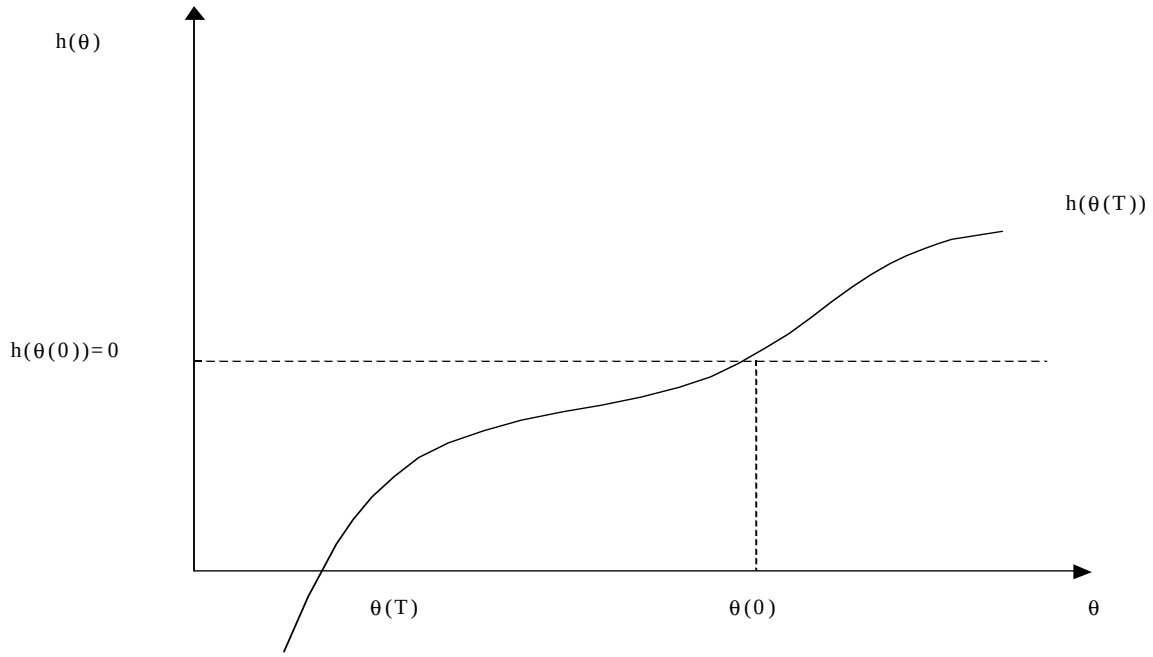


Figure 5.1: Asymptotic normality

Example 5.12

1) Suppose $P[y_t > \theta] = .5\ \forall t$, *where* θ *is the median and let* $g(y_t, \theta) = \begin{cases} 1 & if \quad y_t > \theta \\ -1 & if \quad y_t < \theta \\ 0 & if \quad y_t = \theta \end{cases}$

If $g_T(\theta) = \sum_{t=1}^{T} g(y_t, \theta)$, *then* $E(g(\theta)) = 0$ *and* $g_T(\theta_T) = 0$. *However the g function has a discrete jumps. Since the discontinuity does not get smaller as* $T \to \infty$, *information about* $h_T(\theta)$ *can not be transformed in information about* θ.

2) Consider the problem of minimizing $\frac{1}{T}\sum g(\theta_t)$ *reducing the effect of outliers, i.e., minimizing a quadratic function up to* $\bar{\theta}$ *and an absolute function afterwards. The resulting estimator has jumping derivatives. Therefore, condition (ii) is not satisfied.*

To insure that pathologies like those in example 5.12 do not occur, we use the following:

Definition 5.2 *(continuity in the mean) The function* $f(y_t, \theta)$ *is continuous in the mean at* θ_0 *if and only if there exists a function* $\phi(y_t, \varepsilon)$ *such that* $\max ||f(y_t, \theta) - f(y_t, \theta_0)|| < \phi(y_t, \varepsilon),\ \forall ||\theta - \theta_0|| < \epsilon$ *and* $E[\phi(y_t, \varepsilon)] = \varphi_\infty(y_t, \epsilon)$ *satisfies* $\phi_\infty(y_t, \varepsilon) \longrightarrow 0$ *as* $\varepsilon \to 0$.

Note that continuity in the mean is stronger than simple continuity since the choice of (ϕ, ε) has to work for all y and θ. When applied to example 5.12 it insures that, around θ_0, the size of the maximum jump in h_T has finite expectations.

Example 5.13 *In figure 5.2, $h_T(\theta)$ is continuous, $Eh(\theta)$ is finite since there is only a finite number of spikes but* $\max[h(\theta)]$ *may be infinite if any spike has infinite height. Hence, continuity of derivative and finiteness of expectations do not imply continuity in the mean.*

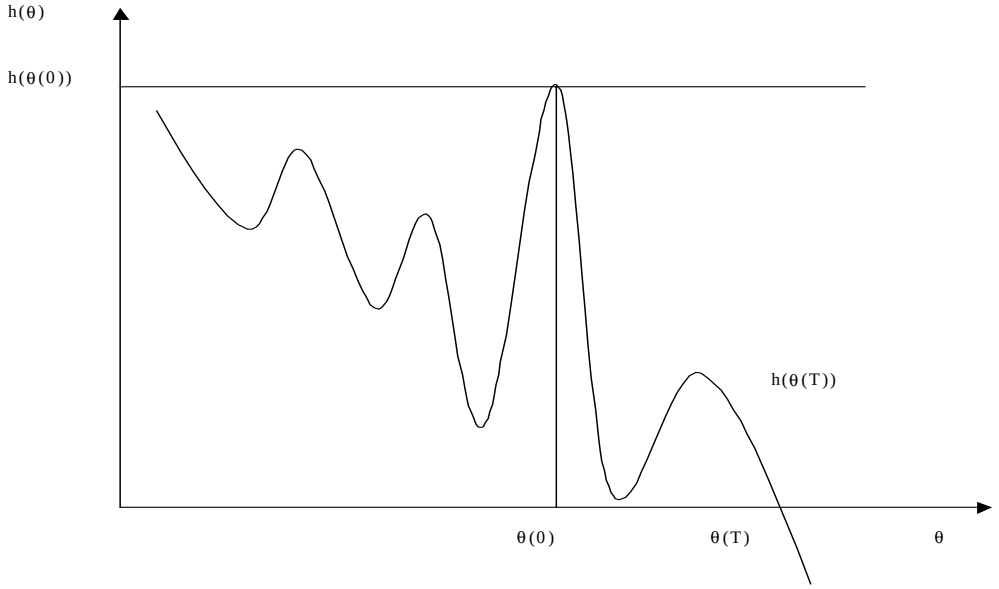


Figure 5.2: Convergence in the mean

Exercise 5.21 *Show the asymptotic covariance matrix of the NLLS estimator using the formula* $H((\theta_0)'\Sigma_h^{-1}H(\theta_0))^{-1}$. *Show that for a ML estimator* $\Sigma_\theta = -H(\theta_0)^{-1}$.

5.3.2 Estimating the Covariance Matrix

We have seen that $\Sigma_\theta = (H(\theta_0)'\Sigma_h^{-1}H(\theta_0))^{-1}$ when $W = \Sigma_h^{-1}$. Estimates of $H(\theta)$ and Σ_h can be obtained using $\{\frac{1}{T}\sum_t[\frac{\partial h_T(\theta_T)}{\partial \theta_T'}]\}$ and $\frac{1}{T}\sum_{t=1}^{T} h_t h_t'$. When g_t is not a martingale difference, the central limit theorem for serially correlated processes can be used to show that the asymptotic covariance matrix is $\Sigma_\theta^+ = (H(\theta_0)'(\Sigma_h^+)^{-1}(H(\theta_0)')^{-1}$ when $W = (\Sigma_h^+)^{-1}$. Here Σ_h^+ is the frequency zero of the spectrum of h_t and can be estimated using $\hat{\Sigma}^+ = \sum_{\tau=-\infty}^{\infty}\sum_t h_t h_{t-\tau}'$. Three assumptions are needed for this result to hold (see chapter 1): (i) $E(y_t y_t')$ exists and is finite, (ii) $E(y_t|y_{t-\tau}, y_{t-\tau-1}, \ldots) \overset{q.m.}{\to} 0$, (iii) $Rev_{t-\tau}(t) = E(y_t|y_{t-\tau}, y_{t-\tau-1}, \ldots) - E(y_t|y_{t-\tau-1}, y_{t-\tau-2}, \ldots)$ is such that $\sum_j (E(Rev_{t-\tau}(t)Rev_{t-\tau}(t)'))^{0.5}$ exists and it is finite.

When deviations from the martingale assumption are of known form, one should use this information in constructing an estimate of Σ_h^+. The next exercise examines situations where g_t is linear covering the cases studied by Hansen and Hodrick (1980), Hansen and Singleton (1982), Cumby, Obstfeld and Huizinga (1983) and Hansen and Sargent (1982).

Exercise 5.22 *Let* $g_t = [e_t \otimes z_t]$ *where* $\otimes$ *is the Kroneker product,* e_t *is a vector of residuals and* z_t *a vector (matrix) of instruments and let* $g_\infty = 0$
i) (Serial correlation up to lag τ *and conditional homoskedasticity) Suppose that* $E_t[e_t|\ z_t, e_{t-\tau}, z_t, e_{t-\tau-1}\cdots] = 0$. *Show that* $E_t[g_t|g_{t-\tau}, g_{t-\tau-1}\ldots] = 0$. *Let* $E[e_t e_{t-\tau}|z_t, e_{t-\tau}, z_{t-\tau}\cdots] = ACF_e(\tau)$. *Show that* $\Sigma_h^+ = \sum_{i=-\tau+1}^{\tau-1} ACF_e(i) \otimes ACF_z(i)$.
ii) (Serial correlation up to lag τ *and conditional heteroskedasticity) Let* $E_t[e_t|z_t, e_{t-\tau}, z_t, e_{t-\tau-1}, \ldots] = 0$ *and* $E[g_t g_{t-\tau}'] = ACF_{ez}(\tau)$. *Show that* $\Sigma_h^+ = \sum_{i=-\tau+1}^{\tau-1} ACF_{ez}(i)$. *Show the form of a typical element of* $ACF_{ez}(i)$.

As we have seen there are cases when the residuals of Euler equations of a DSGE model display serial correlation of known form. Two were described in section 2; two more are considered next.

Exercise 5.23 *(Eichenbaum, Hansen, Singleton) Suppose a representative agent ranks consumption and leisure streams according to* $E\sum_t \beta^t \frac{(c_t^*)^\vartheta((1-N_t)^*)^{1-\vartheta})^{1-\varphi}-1}{1-\varphi}$ *where* $c_t^* = (1+\gamma_1\ell)c_t$ *and* $(1-N_t)^* = (1+\gamma_2\ell)(1-N_t)$ *subject to the constraint* $c_t + sa_{t+1} = w_t N_t + (1+r_t)sa_t$ *where* sa_t *are savings and* $w_t N_t$ *is labor income.*
i) Show that the marginal utility of consumption and leisure satisfy $U_{c,t} = E_t(1+\beta\gamma_1\ell^{-1})U_{c^*,t}$; $U_{N,t} = E_t(1+\beta\gamma_2\ell^{-1})U_{N^*,t}$.
ii) Show that the optimal intratemporal allocation is $E_t\{w_t[(1+\beta\gamma_1\ell^{-1})\vartheta((1+\gamma_1\ell)c_t)^{\vartheta(1-\varphi)-1}((1+\gamma_2\ell)(1-N_t))^{(1-\vartheta)(1-\varphi)}+((1+\gamma_1\ell)c_t)^{(1-\varphi)\vartheta}(1+\beta\gamma_2\ell^{-1})(1-\vartheta)((1+\gamma_2\ell)(1-N_t))^{(1-\varphi)(1-\vartheta)-1}]\} = 0$ *which collapses to the standard condition that the marginal rate of substitution equals the real wage if* $\gamma_1 = \gamma_2 = 0$
iii) Show that the Euler equation for saving accumulation is $E\{r_{t+1}[\beta(1+\beta\gamma_1\ell^{-1})\ ((1+\gamma_1\ell)c_{t+1})^{\vartheta(1-\varphi)-1}((1+\gamma_2\ell)(1-N_{t+1}))^{(1-\vartheta)(1-\varphi)}] - (1+\beta\gamma_1\ell^{-1})((1+\gamma_1\ell)c_t)^{\vartheta(1-\varphi)-1}((1+\gamma_2\ell)(1-N_t))^{(1-\vartheta)(1-\varphi)}\} = 0$. *Show that if the relationships in ii) or in iii) are used to estimate the unknown parameters, the* g_t *function is not a martingale difference. Show exactly the serial correlation structure present. Construct a covariance matrix for the GMM estimator which takes this correlation structure into account.*

Exercise 5.24 *(West and Wilcox) Consider a (partial equilibrium) setup where firms maximize the expected discounted value of future cash flows and have a cost function which includes linear and quadratic costs of producing and changing inventories. Let* Sl_t *be sales,* GDP_t *production,* Iv_t *end of the period inventories,* rc_t *total costs,* p_t *the price,* e_t *a cost shock, observable to the firm but unobservable to the econometrician and* β *a discount factor. The objective function is* $E_t\sum_t \beta^t(p_t Sl_t - rc_t)$ *where* $GDP_t = Sl_t + Iv_t - Iv_{t-1}$, rc_t *are proportional to* $0.5b_0\Delta GDP_t^2 + 0.5b_1 GDP_t^2 + 0.5b_2(Iv_{t-1} - b_3 Sl_t)^2 + Iv_t e_t$ *where* $b_j,\ j = 0,1,2,3$ *are parameters and* e_t *a shock. Show that the Euler condition is* $E_t(b_0(\Delta GDP_t - 2\beta\Delta GDP_{t+1} + \beta^2\Delta GDP_{t+2}) + b_1(GDP_t - \beta GDP_{t+1}) + \beta b_2(Iv_t - b_3 Sl_{t+1}) + \varrho + e_t) = 0$ *where* ϱ *is a constant. Argue that it displays MA components of order 2. Describe what instruments you would use to estimate* b_j *and how to construct their asymptotic covariance matrix.*

In general, the presence of adjustment costs, time nonseparability in the utility function;

multiperiod forecasts or time aggregation (see Hansen and Singleton (1988)) may produce orthogonality conditions which display serial correlation of known form.

When serial correlation is of unknown form and T is finite, one is forced to estimate Σ_h^+ truncating the infinite sum, that is, one uses $\Sigma_T^+ = ACF(0) + \sum_{i=1}^{J(T)} [ACF(i) + ACF(i)'] = \sum_{i=-\infty}^{\infty} \mathcal{K}(i, J(T)) ACF(i)$ where $J(T)$ is a function of T controlling the number of covariances included and $\mathcal{K}(i, J(T))$ is the Box-Car Kernel (see Chapter 1).

The truncation clearly biases Σ_T^+, but the bias vanishes reasonably fast as $T \to \infty$ (see Priestley (1980, p. 458)). Unfortunately, for arbitrary $J(T)$, Σ_T^+ need not be positive semi-definite. Newey and West (1987) propose to use the Bartlett kernel , $\mathcal{K}(i, J(T)) = 1 - \frac{i}{J(T)+1}$ in place of the Box-Car kernel since it insures that Σ_T^+ is positive semi-definite. This kernel also truncates after $J(T)$ but reduces the importance of included elements using weights that decline with i. Since Σ_h^+ gives unitary weights to all ACF elements, this kernel induces an additional source of bias, which typically dominates the one induced by truncation.

In general, the properties of Σ_T^+ depend on the way $J(T)$ is chosen. With the Bartlett kernel if $J(T) \to \infty$ as $T \to \infty$, and $J(T)/T^{1/3} \to 0$, then $\Sigma_T^+ \xrightarrow{P} \Sigma$. That is, the biases introduced by the Bartlett kernel reduce the convergence rate of Σ_T^+ from $T^{1/2}$ to $T^{1/3}$. This means that, for the sample sizes used in macroeconomics, covariance estimates computed with this kernel may be far away from the true Σ_h. Since $W = \Sigma_h^{-1}$, poor small sample estimates of Σ_T^+ may result in poor small sample properties of optimal GMM estimators. Furthermore, if Σ_T^+ has a large bias or a large MSE in finite samples, inference may be severely distorted. For example, t-tests may over reject the null.

Nowadays, it has become common to construct Heteroskedasticy and Autocorrelation consistent (HAC) covariance matrices. HAC estimates are typically of two types: kernel based (non-parametric) and parametric based. In both cases a number of choices, which may influence the properties of estimates, must be made.

Algorithm 5.2 *(Kernel based HAC)*

1) *Obtain an estimate of the orthogonality conditions filtering out some serial correlation.*

2) *Given a kernel, choose the bandwidth parameter $J(T)$ optimally or use rules of thumb.*

3) *Provide estimates of unknown optimal quantities in 2)*

4) *Calculate the spectral density of the orthogonality conditions obtained in 1).*

5) *Calculate HAC consistent estimates of the original orthogonality conditions*

When the orthogonality conditions display autocorrelation of unknown form it is a good idea to eliminate part of this correlation before Σ_T^+ is calculated. The reason is simple. When $h_T(\theta)$ is serially correlated it will have a non-flat spectrum. Kernel estimators average the spectrum of $h_T(\theta)$ over an interval of frequencies. Priestley (1980, p.458) showed that if

one estimates a function $f(\theta)$ at θ_0, averaging it at a number of points in a neighborhood of θ_0, then the estimator is unbiased only if $f(\theta)$ is flat over the neighborhood. Otherwise, the bias depends on the degree of non-constancy of $f(\theta)$. Hence, if we filter $h_T(\theta)$, so that it has a flatter spectrum in the required interval, a kernel-based estimator will have much better properties. Since the purpose of filtering is not to whiten h_t but only to reduce serial correlation, a researcher regresses h_t on an arbitrary number of lags (typically, one).

We presented several types of kernel estimators in chapter 1. The HAC literature has concentrated primarily on three: the Bartlett kernel, the Parzen Kernel (see Gallant (1987)) and the quadratic spectral (QS) kernel (see Andrews (1991)).

In all cases the choice of $J(T)$ is crucial. To choose this parameter, one can use rough rules of thumb, e.g. $J(T) = T^{\frac{1}{3}}$ or, given a kernel, one can choose it optimally, requiring that Σ_T^+ is positive definite. It turns out that selecting $J(T)$ is equivalent to choosing $J_2(w_0)$ in $J(T) = J_1(w_0)[J_2(w_0)T]^{\frac{1}{2w_0+1}}$ where w_0 is the rate of convergence of the kernel (which is equal 2 for Parzen and QS kernels and 1 for the Bartlett kernel). Since Σ_T^+ is consistent but asymptotically biased, optimization requires choosing $J_2(w_0)$ to minimize the MSE of $\mathcal{W}'(\Sigma_T^+ - \Sigma^+)\mathcal{W}$, given a weighting matrix $\mathcal{W}$. Because of this bias, the resulting estimator is not asymptotically efficient. When $\mathcal{W}$ is diagonal and $\mathcal{W}_{ii} = vec(WW')_{ii}$, the optimal bandwidth parameter is $J_2(w_0) = [\frac{W'(\Sigma^{w_0})W}{W'\Sigma W}]^2$. (see Den Haan and Levine (1996)) where $\Sigma^{(w_0)} = \sum_\tau |\tau|^{w_0} \sum_t h_t h_{t-\tau}$ is the w_0-th derivative of Σ_T^+ and measures its smoothness around frequency 0, and $J_1(w_0) = 1.1447$ if $w_0 = 1$ and $J_1(w_0) = 1.3221$ if $w_0 = 2$.

It turns out that among the kernels which generate positive semi-definite estimators, the QS is optimal (Andrews (1991)). Two features of this kernel should be emphasized: first it does not truncate the ACF of h_t. Second, since the weights it gives to elements within $\pm J(T)$ are larger than those given by, e.g. the Bartlett kernel, the second source of bias is also reduced. This bias reduction, has important implications: Σ_T^+ now converges to Σ^+ faster than the Bartlett Kernel ($T^{\frac{2}{5}}$ vs. $T^{\frac{1}{3}}$). Simulation experiments however suggest that the small sample performance of the two kernels are roughly similar and that the choice of $J(T)$ is what matters most.

The third step involves providing estimates of Σ_0^w and Σ (the optimal choice of $J_2(w_0)$ is non-operative). There are two approaches in the literature: Andrews and Mohanan (1992) estimate AR(1) representations for the filtered orthogonality conditions and use these estimates to obtain an estimate of $J_2(w_0)$. In this case $\hat{J}_2(w_0) = \frac{\sum_i W_i 4\hat{\rho}_i^2 \hat{\sigma}_i^4 (1-\hat{\rho}_i)^{-6}(1+\hat{\rho})^{-2}}{\sum_i W_i \hat{\sigma}_i^4 (1-\hat{\rho}_i)^{-4})}$ for $w_0 = 1$ and $\hat{J}_2(w_0) = \frac{\sum_j w_j 4\hat{\rho}_j^2 \hat{\sigma}_j^4 (1-\hat{\rho}_j)^{-8}}{\sum_j w_j \hat{\sigma}_j^4 (1-\hat{\rho}_j)^{-4}}$ for $w_0 = 2$ where $\hat{\rho}_j(\hat{\sigma}_j)$ is an estimate of the AR(1) coefficient (standard deviation of the error) for the filtered condition j. Newey and West (1994) instead choose an automatic procedure which depends on T and on one (arbitrary) parameter. Here $\hat{\Sigma}_0^w = \sum_{\tau=-J(T)}^{J(T)} |\tau|^{w_0} \frac{1}{T} \sum_t e_t e'_{t-i}$ where e_t are the filtered orthogonality conditions obtained in i); $J(T) = b_1(0.01T)^{2/9}$ for $w_0 = 1$; $J(T) = b_2(0.01T)^{2/25}$ for $w_0 = 2$; $b_1 = 4$ or 12 and $b_2 = 3$ or 4.

Once an optimal kernel is obtained (call it $\mathcal{K}^\dagger(i, J(T))$, an estimate of the covariance matrix of the filtered error e_t is $\Sigma_e^\dagger = \sum_i \mathcal{K}^\dagger(i, J(T)) \frac{1}{T} \sum_t e_t e'_{t-i}$ and an estimate of the

covariance matrix of the original conditions is $\Sigma_T^\dagger = [I_N - \sum_i A_i]^{-1}\Sigma_e^\dagger([I_N - \sum_i A_i]^{-1})'$ where A_i are the i-th AR coefficients obtained in step 1).

There are two important features of kernel based HAC estimates worth mentioning. First, J(T) must grow with the sample size for the resulting estimator to be consistent. This is unfortunate since it forces the bandwidth parameter to grow with T even when the serial correlation of h_t is known to be finite. Second, even optimal kernel estimators converge slowly. Hence, they may have worse small sample properties than a parametric estimator (which converges at $\sqrt{T}$ rate).

Algorithm 5.3 *(Parametric HAC)*

1) *For each j, specify a VAR for h_{jt} and select the order of the VAR optimally.*

2) *Calculate the spectral density of the "prewhiten" orthogonality conditions.*

3) *Calculate HAC consistent estimates of the original orthogonality conditions.*

In step i) one specifies the autoregression $h_{jt} = \sum_{i=1}\sum_{\tau=1} A_{i\tau}h_{it-\tau} + e_{jt}$ and chooses the lag length using information criteria (see Chapter 4). Note that the same number of lags of each h_{it} enter the autoregression of h_{jt}. Den Haan and Levin (1996) show that starting the search process from $\bar{\tau} = T^{\frac{1}{3}}$ produces consistent estimates. Once white noise residuals are obtained an estimate of Σ_e is $\Sigma_e^+ = \frac{1}{T}\sum_t e_t e_t'$ and $\Sigma_T^+ = [I_N - \sum A_\tau]^{-1}\Sigma_e^+([I_N - \sum A_\tau]^{-1})'$.

Note that in the parametric approach, the question of positive definiteness does not arise since $\Sigma_T^\dagger$ is positive definite by construction. Also, because the parametric estimator produces smaller biases than kernel estimators, it has better convergence properties.

Example 5.14 *Suppose that h_t has two components. Then prewhitening works as follows. For each $i = 1, 2$ determine the lag length of $A_{i1}(\ell)$ and $A_{i2}(\ell)$ in $h_{it} = A_{i1}(\ell)h_{it-1} + A_{i2}(\ell)h_{i't-1} + e_{it}, i \neq i'$ which could be different for different i. Collect the two equations into a VAR and transform it into a companion form $\mathsf{Y}_t = \mathsf{A}\mathsf{Y}_{t-1} + \mathsf{E}_t$ where $\mathsf{Y}_t = [h_{1t}, h_{2t}]'$. Then var$(\mathsf{E}_t) = \Sigma_E$, var$(\mathsf{Y}_t) = (I - \mathsf{A})^{-1}\Sigma_E((I - \mathsf{A})^{-1})'$ and cov$(\mathsf{Y}_t, \mathsf{Y}_{t-\tau}) = \mathsf{A}^\tau$ var(Y_t).*

5.3.3 Optimizing the Asymptotic covariance matrix

There are a number of ways to make GMM estimators efficient. For example, one can choose W to minimize the asymptotic covariance matrix of θ. As in the linear framework, it is optimal to set $W = (E(h_t h_t'))^{-1}$ if h_t is a martingale difference - with the obvious adjustments if serial correlation is present.

As mentioned, DSGE models typically deliver orthogonality conditions of the form $E[e(y_t, \theta)|z_t] = 0$ where z_t is a set of instruments in agent's information set and e_t are the residuals of an Euler equation. This restriction implies $E[z_t' e(y_t, \theta)] = 0$ but also that $E[f(z_t)' e(y_t, \theta)] = 0$ for any measurable function f. What is the optimal f?

Let $g(y_t, \theta) = f(z_t)' e(y_t, \theta)$ and assume $g_\infty(\theta) = 0$. If y_t, z_t are jointly stationary and ergodic; $E[e(y_t, \theta)|z_t, z_{t-1}, \ldots, e_{t-1}, e_{t-2}\cdots] = 0$; $E[e(y_t, \theta)^2|z_t, z_{t-1}, \ldots, e_{t-1}, e_{t-2}, \ldots]$

$= \sigma_e^2$; if $H \equiv E\left|\frac{\partial h(y_t,\theta_0)}{\partial \theta}\right| = E\left|\frac{f(z_t)'\partial e(y_t,\theta_0)}{\partial \theta}\right|$ and $\Sigma \equiv E[h(y_t,\theta_0)h(y_t,\theta_0)'] = \sigma_e^2 E[f(z_t)'f(z_t)]$, then $var[\sqrt{T}(\theta_T - \theta_0)] = (H\Sigma^{-1}H')^{-1} = (E[\frac{\partial e_t}{\partial \theta}' f(z_t)]^{-1} E[f(z_t)'f(z_t)] E[f(z_t)'\frac{\partial e_t}{\partial \theta}])^{-1}\sigma_e^2$.

The expression in parenthesis is the inverse of the population covariance matrix of the predicted values of the linear regression of $\frac{\partial e_t}{\partial \theta}$ on $f(z_t)$. Therefore, to minimize $\text{var}(\theta_T)$, one should select f to maximize the correlation between $f(z_t)$ and $\partial e_t/\partial\theta$.

Exercise 5.25 *Show that it is optimal to set $f(z_t) = E[\frac{\partial e_t}{\partial \theta}(y_t,\theta_0)|z_t]$ (Hint: Any other $\tilde{f}(z_t)$ will produce a covariance matrix $\tilde{\Sigma}_\theta$ such that $\tilde{\Sigma}_\theta - \Sigma_\theta$ is positive semi-definite.)*

Intuitively the result of exercise 5.25 obtains because the best MSE predictor of a sequence of random variables is its conditional expectation. There are few features of the result that need to be emphasized. First, the optimal $f(z_t)$ is non-unique up to a nonsingular linear transformation of the relationship, i.e. if $f(z)$ achieves the bound also $bf(z)$ will do it where b is a matrix of constants. Second, the formula is non-operative since both e and θ_0 are unknown. Hence a consistent θ_T is typically used to calculate the derivative of e. Third, if $e(y_t,\theta_0)$ is independent of z_t, then for every pair of continuous and measurable functions f_1, f_2 such that $E[f_1(e(y_t,\theta_0))] = 0$, $\tilde{g} = f_2(z_t)'f_1(e(y_t,\theta_0))$ is a potential choice of g function.

Exercise 5.26 *Find the optimal (f_1, f_2) pair in the above problem.*

Exercise 5.27 *Let $y_t^{\theta_0} = \theta_1 + \theta_2 x_t + e_t$. Assume $E[f(z_t)'e_t] = 0$, $f(z_t) = E[\frac{\partial e}{\partial \theta'}|z_t]$, $\theta = [\theta_0, \theta_1, \theta_2]$.*
(i) Show that a NLIV estimator obtained using $f(z_t)$ differs from a NLLS estimator obtained using $f(x_t)$ as the instruments.
(ii) Show that the NLLS estimator is also different from a ML estimator.
(iii) Show that the NLLS estimator is inconsistent. (Hint: NLLS solves $E[f(x)'e_t] = 0$.)

Often in DSGE models the density $f(y|\theta)$ of the endogenous variables can not be calculated. However, one can simulate a sequence y_t and compute approximations to the moments of $f(y|\theta)$ using a law of large numbers (see section 5) . Hence, there are cases where ML is not applicable but GMM is. Since θ_{ML} typically has the smallest asymptotic covariance matrix in the class of estimators which are consistent and asymptotically normal, one may want to know whether it is possible to construct a θ_T which is as efficient as θ_{ML}. In other words, among all possible orthogonality conditions, which ones have the most information for the parameters? Gallant and Tauchen (1996) show that such orthogonality conditions are the scores of each observation.

Example 5.15 *Let y_t be iid with known density $f(y_t,\theta)$. Let θ_T be the GMM estimator associated with orthogonality conditions of the form $E(g_t(\theta, y_t)) = 0$ and suppose Σ_θ is its asymptotic covariance matrix. Then $\Sigma_\theta \geq -T^{-1}E(\frac{\partial^2 log f(y_t,\theta)}{\partial\theta\partial\theta'}]^{-1}$ with equality holding if $g_t(\theta) = \frac{\partial log f(y_t,\theta)}{\partial\theta}$. Hence the most efficient GMM estimator solves $\frac{1}{T}\sum_t \frac{\partial log f(y_t,\theta)}{\partial\theta'} = 0$.*

Exercise 5.28 *Suppose that y_t is serially correlated. How would you modify the argument of example 5.15 to fit this case?*

5.3.4 Sequential GMM Estimation

There are many applied situations where θ can be naturally separated in two blocks, $\theta = (\theta_1, \theta_2)$, and one may consider sequential estimation of the two sets of parameters or estimation of θ_1, conditional on θ_2. Two cases where this may occur are presented next.

Example 5.16 *Consider ML estimation of the parameters of $y_t = x_t\theta + e_t, e_t \sim (0, \Sigma_e)$. One common procedure is to get a consistent estimate for Σ_e, concentrate the likelihood, and then estimate θ. It turns out that unless x_t is strictly exogenous, the standard formula for Σ_θ obtained using a sequential approach is incorrect - we need to take into account the correlation of x_t and e_t.*

Example 5.17 *Consider the Euler equation for capital accumulation (5.3). Here β, δ and the marginal product of capital f_K are unknown. Since β is typically hard to estimate, one may want to fix it and use (5.3) to get GMM estimates of the real return to capital $f_K + (1-\delta)$. Alternatively, one could use extraneous information to get an estimate of f_K, and use (5.3) to estimate δ and β. In both cases, taking first stage estimates as if they were the true ones, may distort asymptotic standard errors.*

Using a GMM approach it is easy to see what kind of adjustments need to be made. Let $g_T(\theta_1, \theta_2) = \frac{1}{T}\sum_{t=1}^{T} g(y_t, \theta_1, \theta_2)$ and let $g_\infty(\theta_1, \theta_2) = 0$. When θ_2 is known and equal to θ_{20}, the asymptotic distribution of θ_{1T} is obtained using the results of section 3.1.

Exercise 5.29 *Suppose $\theta_2 = \theta_{20}$ is known and that i) $g(y_t, \theta_{10}, \theta_{20})$ is a martingale difference process, $g(y_t, \cdot, \cdot)$ is continuous in the second and third argument, where θ_{10} is a $k_1 \times 1$ vector; ii) y_t is stationary and ergodic; iii) $E[g(y_t, \theta_{10}, \theta_{20})g(y_t, \theta_{10}, \theta_{20})'] = \Sigma < \infty$. Show that $\sqrt{T}g_T(y_T, \theta_{1T}, \theta_{20}) \xrightarrow{D} \mathsf{N}(0, \Sigma_1)$.*

When both θ_{10} and θ_{20} are unknown, let i), ii), iii) of exercise 5.29 be satisfied and assume: iv) $\theta_{1T} \xrightarrow{a.s.} \theta_{10}$, $\theta_{2T} \xrightarrow{a.s.} \theta_{20}$; v) $\left[\frac{\partial h}{\partial \theta_1}, \frac{\partial h}{\partial \theta_2}\right]$ are continuous in the mean at θ_{10}, θ_{20}; vi) $E\left[\frac{\partial h(y_t, \theta_{10}, \theta_{20})}{\partial \theta_1'}, \frac{\partial h(y_t, \theta_{10}, \theta_{20})}{\partial \theta_2'}\right] = [H_{10}, \ H_{20}]$ $|H_{10}| \neq 0$ where $H_{10} and H_{20}$ are $n \times k_1$ and $n \times k_2$ matrices; vii) $\sqrt{T}(\theta_{2T} - \theta_{20}) \xrightarrow{D} \mathsf{N}(0, \Sigma_2)$.

Exercise 5.30 *Show that under i)-vii) $\sqrt{T}(\theta_{1T} - \theta_{10}) \xrightarrow{D} \mathsf{N}(0, (H_{10}')^{-1}\Sigma_1 H_{10}^{-1} + (H_{10}')^{-1}H_{20}'$ $\Sigma_2 H_{20} H_{10}^{-1})$. (Hint: Take Taylor expansion of $h_T(y_T, \theta_{1T}, \theta_{2T})$ around $(\theta_{10}, \theta_{20})$ and make sure all the quantities converge to non-stochastic matrices in the limit).*

Exercise 5.30 shows that the asymptotic covariance matrix of θ_{1T} needs to be adjusted when θ_2 has been estimated. There is one situation when this adjustment is not needed: if the marginal distribution of θ_1 does not depend on θ_2, $H_{20} = 0$, and $\sqrt{T}(\theta_{1T} - \theta_{10}) \xrightarrow{D}$ $\mathsf{N}(0, (H_{10}')^{-1}\Sigma_1 H_{10}^{-1})$. Hence, sequential estimation does not distort standard errors of θ_{1T}.

One special case of the setup we have considered is obtained when θ_{2T} has a degenerate asymptotic distribution and the g function has a special structure.

Exercise 5.31 *Suppose $g_\infty = 0$ and $g(y_t, \theta_1, \theta_2) = \theta_2 g^*(y_t, \theta_1)$ where g^* is a martingale difference process and θ_2 a scalar. Assume that $\theta_{2T} g^*(\theta_{1T}) = 0$ and that $\sqrt{T}(\theta_{2T} - \theta_{20})$ converges in distribution to a constant. Show that $\sqrt{T}(\theta_{1T} - \theta_{10}) \xrightarrow{D} \mathsf{N}(0, ((\theta_{20} H^*)')^{-1} \theta_{20} \Sigma_1^* \theta_{20}' (\theta_{20} H^*)^{-1})$ where $\Sigma_1^* = E[g^*(y_t, \theta_{10}) g^*(y_t, \theta_{10})']$, $H^* = E[\frac{\partial g^*}{\partial \theta_1'}(\theta_{10})]$. Show what happens if $g(y_t, \theta_1, \theta_2) = f(\theta_2) g^\dagger(y_t, \theta_1)$ and f is a continuous (deterministic) function.*

One simple case where the machinery of this section is applicable is the following

Example 5.18 *The literature has suggested two ways of estimating (ρ, θ) in the model*

$$y_t = x_t \theta + e_t \tag{5.22}$$
$$e_t = \rho e_{t-1} + \epsilon_t \quad \epsilon_t \sim (0, \sigma^2) \tag{5.23}$$

1) Use Generalized Least Square to get $\theta_{GLS} = (x' \Sigma_e^{-1} x)^{-1} (x' \Sigma_e^{-1} y)$ where $\Sigma_e = \frac{\sigma^2}{(1-\rho)^2} I$.

2) Estimate ρ_{ols} in (5.23) where $e_{ols,t} = y_t - x_t \theta_{OLS}$. Transform (5.22) to $y_t - \rho_{ols} y_{t-1} = (x_t - \rho_{ols} x_{t-1})\theta + \epsilon_t$ and apply OLS to get $\theta_{2step} = (X'X)'(X'Y)$; $var(\theta_{2step}) = (X'X)^{-1}\sigma^2$ where $X_t = x_t - \rho_{ols} x_{t-1}$, $Y_t = y_t - \rho_{ols} y_{t-1}$ and $X = (X_1, \ldots, X_T)'$, $Y = (Y_1, \ldots, Y_T)'$.

Clearly 2) does not take into account the fact that ρ has been estimated. Properly reading off the marginal asymptotic distribution from the joint one in 1) will account for this. The second approach gives the correct variance for θ only if the asymptotic covariance matrix of θ and ρ is diagonal. In this case, we can treat ρ_{ols} as fixed in the estimation of θ. Problems similar to this emerge in mixed calibration-estimation setups discussed in chapter 7 or in certain panel models (see chapter 8).

5.3.5 Properties of Two-Step Estimators

GMM estimators require iterative procedures and this may make them computationally demanding when h_t is highly non-linear and k is large. As mentioned one could use a two step approach to compute an approximation to the full GMM estimator. Under what conditions a two step GMM estimator will asymptotically be the same as a fully iterative GMM estimator? It turns out that if the initial estimator θ_T^1 is a $\sqrt{T}$-consistent estimator and it is bounded in probability, the difference between θ_T^2 and θ_T vanishes asymptotically. While the first condition is difficult to verify, for the second it suffices that $\theta_T^1 \xrightarrow{D} \theta_0$ (see chapter 1), which is easy to check in practice.

In general, one should prefer full iterative GMM to a two-step GMM estimation when the initial estimator has a large covariance matrix and when the space of h_t is not well approximated by a quadratic function. For example, if h_t is highly nonlinear, small changes in θ may cause large changes in H_t. Note also that the sample size needed for consistency of θ_T^2 and of θ_T may be very different: θ_T^1 may be in the tail of the asymptotic distribution and if the objective function is flat, it may take a large T for θ_T^2 to approximate θ_T.

We ask the reader to verify the asymptotic equivalence of θ_T^2 and θ_T in the next exercise

Exercise 5.32 *Suppose y_t is stationary and ergodic ; $g(y_t, \theta_0)$ is a martingale difference; $g_\infty(\theta) = 0$, $E[h(y_t, \theta_0)h(y_t, \theta_0)'] = \Sigma < \infty$, $H_T = \frac{\partial h}{\partial \theta'}(y_t, \theta)$ exists and is continuous in the mean at θ_0; $E[H_T(\theta_0)] = H$, $|H| \neq 0$; $h_T(\theta_T) = 0$; $\theta_T \xrightarrow{P} \theta_0$. Assume that θ_T^1 is a $\sqrt{T}-$consistent estimator of θ_0 such that $\sqrt{T}(\theta_T^1 - \theta_0)$ is bounded in probability and let $\theta_T^2 \equiv \theta_T^1 - \left[\frac{\partial h_T(\bar{\theta}_T)}{\partial \theta}\right]^{-1} h_T(\theta_T^1)$ where $\bar{\theta}_T \in [\theta_T^1, \theta_T^2]$. Show that $\sqrt{T}(\theta_T^2 - \theta_0) \xrightarrow{D}$ $\mathsf{N}(0, (H')^{-1}\Sigma H^{-1})$ that $\sqrt{T}(\theta_T^2 - \theta_T) \xrightarrow{P} 0$, where θ_T is the fully iterative GMM estimator.*

5.3.6 Hypotheses Testing

We are concerned with the general problem of testing whether a vector of (possibly non-linear) restriction of the form $R(\theta) = 0$ holds. The discussion in this subsection is general: we specialize the setup in various examples and exercises. We assume that $g_\infty(\theta) = 0$; that θ_T solves $h_T(\theta_T) = 0$ (so that n=k) and that $\sqrt{T}(\theta_T - \theta_0) \xrightarrow{D} \mathsf{N}(0, \Sigma_\theta)$. Under the null, $R(\theta_0) = 0$.

- Wald type test

To derive a Wald type test note that even though $R(\theta_T) \neq 0$, it should be small with high probability if $\theta_T \xrightarrow{P} \theta_0$. Assume that $R(\theta)$ is a smooth function with, at least, the first derivative and that $\frac{\partial R_i(\bar{\theta})}{\partial \theta_j'}$ is continuous in the mean (and full rank). Taking an exact Taylor expansion of $R(\theta_T)$ around $R(\theta_0)$ we have

$$R(\theta_T) = R(\theta_0) + \mathcal{R}(\bar{\theta})(\theta_T - \theta_0) \tag{5.24}$$

where $\mathcal{R}_{ij}(\bar{\theta}) = \frac{\partial R_i(\bar{\theta})}{\partial \theta_j'}$ $i = 1, 2, \ldots$ and $\bar{\theta} \in (\theta_T, \theta_0)$.

Using continuity of $\frac{\partial R_i(\bar{\theta})}{\partial \theta_j'}$ and consistency of θ_T, we have $\mathcal{R}(\bar{\theta}) \xrightarrow{P} \mathcal{R}(\theta_0)$, elementwise. Then from (5.24) we have that $\sqrt{T}R(\theta_T) = \sqrt{T}\mathcal{R}(\theta_0)(\theta_T - \theta_0) \xrightarrow{D} \mathsf{N}(0, \Sigma_R = \mathcal{R}(\theta_0)\Sigma_\theta\mathcal{R}(\theta_0)')$ by the asymptotic normality of $\theta_T - \theta_0$. Therefore, a test for $R(\theta) = 0$ can be conducted using:

$$Wo = TR(\theta_T)'\Sigma_R^{-1}R(\theta_T) \sim \chi^2(dim(R)) \tag{5.25}$$

(5.25) is entirely based on the local properties of $R(\theta_T)$ around θ_0. When g_T is the score of the log likelihood function, (5.25) is a standard Wald test.

Exercise 5.33 *Find a consistent estimator of $\mathcal{R}(\theta_0)$. What happens to (5.25) if $R(\theta_0) \neq 0$.*

- Lagrange multiplier type test

In some problems the imposition of restrictions makes estimation and testing easier.

Example 5.19 *Suppose you want to estimate the parameters of the following nonlinear model* $\frac{y_t^{\alpha_0}-1}{\alpha_0} = \alpha_1 + \alpha_2 \frac{x_t^{\alpha_0}-1}{\alpha_0} + e_t$. *If* $\alpha_0 = 0$ *the model reduces to* $\ln y_t = \alpha_1 + \alpha_2 \ln x_t + e_t$ *while if* $\alpha_0 = 1$ *it reduces to* $y_t = (\alpha_1 - \alpha_2 + 1) + \alpha_2 x_t + e_t$ *and in both cases estimates of* α_1 *and* α_2 *can be obtained with simple least square techniques. Given these estimates, we may want to test whether* $\alpha_0 = 0$ *or* $\alpha_0 = 1$ *is more probable.*

In cases like those of example 5.19 it may be useful to design a test which uses the local properties of $h_T(\theta)$ and $R(\theta)$ around θ_R, a restricted estimator. Let θ_R solve $h_T(\theta_R) = 0$ and assume that $\sqrt{T}(\theta_R - \theta_0) \xrightarrow{P} 0$. Expanding $h_T(\theta_T)$ and $R(\theta_T)$ around θ_R we have:

$$h_T(\theta_T) = h_T(\theta_R) + \frac{\partial h_T(\bar{\theta})}{\partial \theta'}(\theta_T - \theta_R) \tag{5.26}$$

$$R(\theta_T) = R(\theta_R) + \frac{\partial R(\bar{\theta})}{\partial \theta'}(\theta_T - \theta_R) \tag{5.27}$$

where $\bar{\theta} \in (\theta_T, \theta_R)$. From (5.26) and given assumptions made $\sqrt{T}(\theta_T - \theta_R) = \sqrt{T}(\frac{\partial h_T(\bar{\theta})}{\partial \theta'})^{-1}$ $(h_T(\theta_T) - h_T(\theta_R)) \xrightarrow{D} \mathsf{N}(0, \Sigma_\theta = (H')^{-1}\Sigma_h H^{-1})$.

Exercise 5.34 *Give conditions sufficient to insure the above result. Intuitively explain why the distributions of* $(\theta_T - \theta_0)$ *and* $(\theta_T - \theta_R)$ *are the same.*

Using (5.27) and noting that $R(\theta_R) = 0$ by construction, $\sqrt{T}R(\theta_T) = \frac{\partial R(\bar{\theta})}{\partial \theta'}\sqrt{T}(\theta_T - \theta_R) \xrightarrow{D} \mathsf{N}(0, \Sigma_R = \mathcal{R}(\theta_0)\Sigma_\theta \mathcal{R}(\theta_0)')$. Therefore, a test for the null hypothesis is

$$LM = TR(\theta_T)'\Sigma_R^{-1}R(\theta_T) \sim \chi^2(dim(R)) \tag{5.28}$$

It is easy to verify that the quadratic forms in (5.25) and (5.28) are similar - (5.25) uses the properties of an unrestricted estimator while (5.28) those of a restricted estimator - and asymptotically equivalent. In general, this is not the case (see e.g. Engle (1983)). Here it occurs because restricted and unrestricted estimators have the same asymptotic covariance matrix which, in turn, is due to the fact that θ_R is a $\sqrt{T}$-consistent estimator of θ_0. Note that for the equivalence result to hold W_T must be optimally chosen and the same optimal W must be used in both specifications.

This class of tests based on the local properties of the h_T function around θ_R is called Lagrange multiplier test . When $g_T = \frac{1}{T}\sum \frac{\partial \log f(y_t,\theta)}{\partial \theta}$ and $E(g(y_t, \theta)) = 0$, θ_R is the maximum likelihood estimator obtained subject to $R(\theta) = 0$. Then, (5.28) tests the hypothesis that the Lagrangian multiplier on the restriction is zero (see Judge, et at. (1985, p.182)).

There is an alternative way to check the properties of $h_T(\theta)$ and $R(\theta)$ around θ_R. Let $\mathcal{C}_T \xrightarrow{P} \mathcal{C}_0$ be a $(n - \dim(R)) \times n$ matrix and let $[\mathcal{C}_T h_T(\theta_R), R(\theta_R)]' = 0$ be a $n \times 1$ vector.

Exercise 5.35 *Give conditions under which* $\sqrt{T}(\theta_R - \theta_0) = (diag\ [\frac{\partial h(\bar{\theta})}{\partial \theta'},\ \frac{\partial R(\bar{\theta})}{\partial \theta'}])^{-1} \times$ $\begin{bmatrix} \sqrt{T}\mathcal{C}_T g_T(\theta_0) \\ \sqrt{T}R(\theta_0) \end{bmatrix} \xrightarrow{D} \mathsf{N}(0, \Sigma^\dagger = (diag[\frac{\partial h(\bar{\theta})}{\partial \theta'}, \frac{\partial R(\bar{\theta})}{\partial \theta'}]')^{-1}\Sigma_R(diag[\frac{\partial h(\bar{\theta})}{\partial \theta'}, \frac{\partial R(\bar{\theta})}{\partial \theta'}]^{-1})$ *where* $\Sigma_R =$

$\begin{bmatrix} \mathcal{C}_0\Sigma\mathcal{C}_0' & 0 \\ 0 & 0 \end{bmatrix}$. *(Hint: Expand* $[\mathcal{C}_T h_T(\theta_R), R(\theta_R)]'$ *around* θ_0 *and make sure all quantities converge to the proper limits). Show that a test for the hypothesis that* $R(\theta) = 0$ *is* $LM1 = T[(\theta_R - \theta_0)(\Sigma^\dagger)^{-1}(\theta_R - \theta_0)] \sim \chi^2(n - dim(R))$. *Why are the degrees of freedom of the asymptotic distributions of LM and LM1 different?*

- Distance Test

Distance tests examine whether two estimators (a restricted and an unrestricted one) are close in some metric. They are useful when the minimized value of the criterion function is never directly computed (as e.g. with maximum likelihood). Let θ_R solve $h_T(\theta_R) = 0$ and let θ_T be a $\sqrt{T}-$consistent (unrestricted) estimator. Then $0 = h_T(\theta_R) = h_T(\theta_T) + H(\bar{\theta})(\theta_R - \theta_T)$ with $\bar{\theta} \in (\theta_R, \theta_T)$ where $H_{ij} = \frac{\partial h_{Ti}(\bar{\theta})}{\partial \theta_j}$.

Exercise 5.36 *Give sufficient conditions to insure that* $H(\bar{\theta}) \xrightarrow{p} H$ *and* $\sqrt{T}(\theta_R - \theta_T) \xrightarrow{D} \mathsf{N}(0, \Sigma_\theta)$ *where* $\Sigma_\theta = (H(\theta_0)')^{-1}\Sigma(H(\theta_0))^{-1}$.

Under the conditions of exercise 5.36, a test for the null hypothesis $R(\theta) = 0$ is

$$Dt = T(\theta_R - \theta_T)'\Sigma_\theta^{-1}(\theta_R - \theta_T)' \sim \chi^2(k) \tag{5.29}$$

where Σ =var$(h_T(\theta_T))$. If θ_R is a random vector, then the test has a smaller number of degrees of freedom and this occurs even if $(\theta_T - \theta_R)$ is a $k \times 1$ vector.

Example 5.20 *A likelihood ratio test is a special case of a distance test. Expanding* $\mathcal{L}_T(\theta_R)$ *around a* $\sqrt{T}-$*consistent unrestricted estimator* θ_T *we have* $\mathcal{L}_T(\theta_R) = \mathcal{L}_T(\theta_T) + \frac{\partial \mathcal{L}_T(\bar{\theta})}{\partial \theta'}(\theta_R - \theta_T) + 0.5(\theta_R - \theta_T)'\frac{\partial^2 \mathcal{L}_T(\bar{\theta})}{\partial\theta\partial\theta'}(\theta_R - \theta_T)$. *Since* $\sqrt{T}(\theta_T - \theta_0) \xrightarrow{P} 0$, $\frac{\partial \mathcal{L}_T(\bar{\theta})}{\partial \theta'} \xrightarrow{P} \frac{\partial \mathcal{L}_T(\theta_T)}{\partial \theta'} = 0$. *Hence* $2T(\mathcal{L}_T(\theta_R) - \mathcal{L}_T(\theta_T)) = T(\theta_R - \theta_T)'\frac{\partial \mathcal{L}_T^2(\bar{\theta})}{\partial\theta\partial\theta'}(\theta_R - \theta_T)$. *Then, since* $\frac{\partial^2 \mathcal{L}_T(\bar{\theta})}{\partial\theta\partial\theta'} \xrightarrow{P} \frac{\partial^2 \mathcal{L}_T(\theta_T)}{\partial\theta\partial\theta'} \equiv \Sigma_\theta = (-H^{-1}) = \Sigma_\theta^{-1}$, *we have that* $-2T(\mathcal{L}_T(\theta_R) - \mathcal{L}_T(\theta_T)) \xrightarrow{P} T(\theta_R - \theta_T)'H^{-1}(\theta_R - \theta_T) \sim \chi^2(k - k')$ *and* $k - k'$ *is the number of restrictions.*

- Hausman Test

Hausman's (1978) test is based on the idea that it is not necessary for the unrestricted estimator to be efficient as long as restricted and unrestricted estimators have a joint limiting distribution. Let θ_R be an efficient estimator under the null, i.e. it minimizes the asymptotic covariance matrix; let θ_T be any consistent, not necessarily efficient estimator; let $\theta^0 = [\theta_0, \theta_0]'$, and let $\theta^{TR} = [\theta_T, \theta_R]'$ be such that $\sqrt{T}h_T(\theta^{TR}) \xrightarrow{D} \mathsf{N}(0, \Sigma_\theta)$. If the parameter space is compact, the uniform convergence theorem insures that the asymptotic covariance matrix of $\sqrt{T}(\theta^{TR} - \theta^0)$ has zero off-diagonal elements (so that the two estimators are asymptotically independent). A version of Hausman test is then:

$$Ha = T(\theta_T - \theta_R)'(\Sigma_T - \Sigma_R)^{-1}(\theta_T - \theta_R) \xrightarrow{D} \chi^2(k) \tag{5.30}$$

where Σ_T and Σ_R are the asymptotic covariance matrices of the two estimators. Note that if also θ_T is efficient, $\Sigma_T - \Sigma_R$ is singular. In this case, it is still possible to implement the test by choosing a $k' \times k$ matrix $\mathcal{C}$ such that $|\mathcal{C}(\Sigma_T - \Sigma_R)\mathcal{C}'| \neq 0$ and the test becomes:

$$Ha = T(\theta_T - \theta_R)(\mathcal{C}(\Sigma_T - \Sigma_R)\mathcal{C})^{-1}(\theta_T - \theta_R)' \xrightarrow{D} \chi^2(k') \tag{5.31}$$

Exercise 5.37 *Show the conditions under which (5.29) and (5.30) are equivalent.*

It is important to stress that while several of the tests we consider have the same asymptotic distribution, they may have dramatically different properties in small samples.

Wald tests are easy to implement in practice, as it is shown in the next example.

Example 5.21 *i) Let Σ_{ii} be the i-the element of the optimal Σ and let $R(\theta) = 0$ be $\theta_i = \bar{\theta}_i$. Then $T(\theta_{iT} - \bar{\theta}_i)(\Sigma_{ii})^{-1}(\theta_{iT} - \bar{\theta}_i) \to \chi^2(1)$. Alternatively, $\sqrt{T}\frac{(\theta_{iT} - \bar{\theta}_i)}{\sqrt{\Sigma_{ii}}} \to \mathsf{N}(0,1)$.*
ii) Let the restriction be $R\theta = \bar{\theta}$, then the statistic is $T(R\theta_T - \bar{\theta})'[R\Sigma R']^{-1}(R\theta_T - \bar{\theta}) \xrightarrow{D} \chi^2(k)$

Exercise 5.38 *Consider example 5.10. Provide three test statistics for the hypothesis that the term structure of interest rates is flat in the steady state (i.e. $r_\tau = r_1$) which are robust to the presence of autocorrelation in the orthogonality conditions.*

Example 5.22 *One of the basic assumptions underlying the RBC model of example 5.1 is that agents like to smooth consumption. This implies that the coefficient of relative risk aversion should be positive. One can test this hypothesis in a number of ways. Assuming $u(c) = \frac{c_t^{1-\varphi}}{1-\varphi}$ and using (5.3) we can construct both restricted and unrestricted estimators since the inequality restriction is linear. Wald and Distance statistics are $Wo = T\varphi_T^2/\Sigma_{\varphi_T}$, and $Dt = T(\varphi_T - \varphi_R)^2/\Sigma_{\varphi_T}$ where Σ_{φ_T} (Σ_{φ_R}) is the variance of the unrestricted (restricted) estimator.*

Exercise 5.39 *Consider the asset pricing equations of example 5.3. Suppose that the utility function is logarithmic and that consumption growth has a linear trend. How would you undertake estimation of the discount factor in this case? Can we still use a J-test to examine the validity of the model? How would you test $\beta = 1$?*

Exercise 5.40 *Simulate data for consumption, investment, output and the real interest rate from a RBC model , using $\beta = 0.99$, $\delta = 1$ and $\varphi = 1$ assuming that the log of technology disturbance is AR(1) with persistence 0.9 and standard deviation one. Suppose, as in exercise 5.1, that the utility function display habit persistence in consumption. Construct restricted, unrestricted and minimum distance estimators and test the hypothesis $\gamma = 0$. Repeat the exercise 100 times drawing random technology shocks from a normal distribution. What are the properties of the three tests?*

In some cases we want to test only a subset of the orthogonality conditions. For example, as in exercise 5.10 we may be interested in knowing if income lagged two periods adds explanatory power to our estimates or not. In that case a Hausman test could be used.

Example 5.23 *Continuing with exercise 5.10 let θ_{1T} be the estimator obtained using z_{1T} and θ_2 the estimator obtained using z_2. Clearly θ_{2T} is as efficient as θ_{1T} since it uses more orthogonality conditions. If $var(\theta_{1T} - \theta_{2T}) = var(\theta_{1T}) - var(\theta_{2T})$, Hausman statistic for testing if the second set of orthogonality conditions holds is $(\theta_{1T} - \theta_{2T})(var(\theta_{1T}) - var(\theta_{2T}))^{-1}(\theta_{1T} - \theta_{2T}) \xrightarrow{D} \chi^2(\nu)$ where ν is the minimum between the number of orthogonality conditions tested and the number of conditions minus the number of instruments used for estimation. When $\nu = 1, var(\theta_{1T}) - var(\theta_{2T})$ can be estimated using $T\sigma^2[y'(z_1(z_1'z_1)^{-1}z_1')y - y'(z_2(z_2'z_2)^{-1}z_2')y]$.*

Exercise 5.41 *Consider the situation where agents in one country (say the US) have the option to purchase one period bonds denominated in another currency (say, yen) and let ner_t the (dollar-yen) exchange rate. Show that in equilibrium*

$$0 = E_t[\beta \frac{U_{c,t+1}/p_{t+1}}{U_{c,t}/p_t}((1 + i_{1t}) - \frac{ner_{t+1}}{ner_t}(1 + i_{2t}))] \tag{5.32}$$

where i_{it} is the nominal interest rate on bonds of country i, p_t is the price level and $U_{c,t+1}/p_{t+1}$ the marginal utility of money. Log linearize the condition, assuming $u(c_t) = \ln c_t$. Using nominal balances, nominal interest rates and nominal exchange rate data verify whether (5.32) holds. Test whether agents discount the future or not (i.e. whether $\beta = 1$).

Exercise 5.42 *(McKinlay-Richardson) If a portfolio j of assets is mean variance efficient and if there exists a risk free asset, it must be the case that $E(\tilde{r}_{it}) = \alpha_i E(\tilde{r}_{jt})$ (see exercise 5.2) where $\tilde{r}_{it} = r_{it} - r_{0t}$ is the excess return on asset i at time t, $\tilde{r}_{jt} = r_{jt} - r_{0t}$ is the excess return on a portfolio j at time t, $i = 1, 2, \ldots I$. Derive the orthogonality conditions implied by mean-variance efficiency. Using Euroxx50 stocks data, provide an estimator for α_i which is robust to heteroschedasticity (produced, e.g. if the variance of return i depends on the return on the market portfolio). Test the hypothesis that the efficient frontier holds.*

Example 5.24 *Continuing with example 5.11 we test three hypotheses. First, fixing $\beta = 1$ does not change estimates: a distance style test finds no difference between restricted and unrestricted specifications. Second, we test for full stickiness $\zeta_p = 1$ - this corresponds to excluding marginal costs from the specification (under this null inflation is a simple AR(1) process). A LR test rejects this hypothesis using the output gap in US, UK and the labor share in Germany at the 5% confidence level. Finally, a test of full flexibility $\zeta_p = 0$ is rejected in 5 of the 6 specifications.*

5.4 GMM estimation of DSGE models

The examples considered so far involve the estimation of one equation of a model. At times one may want to examine its full implications so that, e.g., comparison with ML estimates can be performed. In this case systemwide methods are necessary. Typically, there is recursivity in the structure of DSGE models so that estimation can be usefully conducted block by block.

Example 5.25 *Suppose a social planner maximizes* $E_0 \sum_t \beta^t [\frac{c_t^{1-\varphi}}{1-\varphi} + \vartheta_N(1-N_t)]$ *by choices of consumption* (c_t)*, hours* (N_t)*, and capital* (K_{t+1}) *subject to* $G_t + c_t + K_{t+1} = \zeta_t K_t^{1-\eta} N_t^{\eta} + (1-\delta)K_t$ *where* $\ln \zeta_t = \bar{\zeta} + \rho_\zeta \ln \zeta_{t-1} + \epsilon_{1t},\ \epsilon_{1t} \sim (0, \sigma_\zeta^2),\ \ln G_t = \bar{G} + \rho_G \ln G_{t-1} + \epsilon_{4t},\ \epsilon_{4t} \sim (0, \sigma_g^2)$*,* K_0 *given, where . Assume that government expenditure is financed with lump sum taxes or bond creation. The optimality conditions are:*

$$\begin{aligned} \vartheta_N c_t^{\varphi} &= \eta \zeta_t K_t^{1-\eta} N_t^{\eta-1} && (5.33) \\ c_t^{-\varphi} &= E_t \beta c_{t+1}^{-\varphi} [(1-\eta)\zeta_{t+1} K_{t+1}^{-\eta} N_{t+1}^{\eta} + (1-\delta)] && (5.34) \end{aligned}$$

Furthermore, competitive input markets imply that the real wage is $w_t = \eta \zeta_t K_t^{1-\eta} N_t^{\eta-1}$ *and the return to capital is* $r_t = (1-\eta)\zeta_{t+1} K_{t+1}^{-\eta} N_{t+1}^{\eta} + (1-\delta)$*.*

The model has 11 parameters: five structural $\theta_1 = (\beta, \vartheta_N, \varphi, \eta, \delta)$ *and 6 auxiliary ones* $\theta_2 = (\bar{\zeta}, \bar{G}, \rho_\zeta, \rho_g, \sigma_g^2, \sigma_\zeta^2)$*. Hence, we need at least 11 orthogonality conditions to estimate* $\theta = (\theta_1, \theta_2)$*. From the capital accumulation equation and taking unconditional expectations:*

$$E(\delta - 1 + \frac{K_{t+1}}{K_t} - \frac{inv_t}{K_t}) = 0 \tag{5.35}$$

which determines δ *if data on capital and investment are available. The Euler equation (5.34) contains four parameters* $(\beta, \varphi, \eta, \delta)$*. Since* δ *is identified from (5.35) we need to transform (5.34) to produce at least three orthogonality conditions. Since any variable which belongs to the information set at time t can be used as instrument, one could use e.g. a constant, lags of the real retun to capital and of consumption growth to estimate the other three parameters. For example, one could employ*

$$\begin{aligned} E\beta(\frac{c_{t+1}}{c_t})^{-\varphi}[(1-\eta)\zeta_{t+1} K_{t+1}^{-\eta} N_{t+1}^{\eta} + (1-\delta)]) - 1 &= 0 && (5.36) \\ E\{\beta(\frac{c_{t+1}}{c_t})^{-\varphi}[(1-\eta)\zeta_{t+1} K_{t+1}^{-\eta} N_{t+1}^{\eta} + (1-\delta)] - 1\}\frac{c_t}{c_{t-1}} &= 0 && (5.37) \\ E\{\beta(\frac{c_{t+1}}{c_t})^{-\varphi}[(1-\eta)\zeta_{t+1} K_{t+1}^{-\eta} N_{t+1}^{\eta} + (1-\delta)] - 1\} r_{t-1} &= 0 && (5.38) \end{aligned}$$

The intratemporal condition implies

$$E[c_t^{-\varphi} \eta \zeta_t K_t^{1-\eta} N_t^{\eta-1} - \vartheta_N] = 0 \tag{5.39}$$

which also involves three parameters $(\varphi, \eta, \vartheta_N)$*. Given* (φ, η)*, (5.39) determines* ϑ_N*.*

The auxiliary parameters can be estimated using the properties of ϵ_{1t} *and* ϵ_{4t}

$$\begin{aligned} E(\ln \zeta_t - \bar{\zeta} + \rho_\zeta \ln \zeta_{t-1}) &= 0 && (5.40) \\ E(\ln \zeta_t - \bar{\zeta} + \rho_\zeta \ln \zeta_{t-1}) \ln \zeta_{t-1} &= 0 && (5.41) \\ E(\ln \zeta_t - \bar{\zeta} + \rho_\zeta \ln \zeta_{t-1})^2 - \sigma_\zeta^2 &= 0 && (5.42) \\ E(\ln G_t - \bar{G} + \rho_G \ln G_{t-1}) &= 0 && (5.43) \\ E(\ln G_t - \bar{G} + \rho_G \ln G_{t-1}) \ln G_{t-1} &= 0 && (5.44) \\ E(\ln G_t - \bar{G} + \rho_G \ln G_{t-1})^2 - \sigma_g^2 &= 0 && (5.45) \end{aligned}$$

While government expenditure is observable, technological disturbances are not. Therefore an additional auxiliary condition is needed. From the production function and given estimates $\hat{\eta}$ of η we have $\hat{\zeta}_t = \ln y_t - (1-\hat{\eta})\ln K_t - \hat{\eta}\ln N_t$ and $\hat{\zeta}_t$ can be used in (5.40)-(5.42). In sum, the last three conditions could be estimated separately while for the first eight joint or recursive estimation is possible - in the latter case, we need to correct standard errors as described in section 3.4.

Parameter	*Just indentified*	*Over identified*	*Over identified*	*Over identified*
η	*0.18 (0.0002)*	*0.18 (0.0002)*	*0.64*	*0.18 (0.0002)*
φ	*1.0*	*1.0*	*1.0*	*2.0*
δ	*0.0202(0.00022)*	*0.0202 (0.00021)*	*0.0201 (0.00013)*	*0.0208 (0.00013)*
β	*1.007 (0.0005)*	*1.007 (0.0005)*	*0.991 (0.0004)*	*1.012 (0.0009)*
ϑ_N	*3.73 (0.013)*	*3.73 (0.012)*	*2.93 (0.006)*	*0.455 (0.001)*
ρ_ζ	*1.035(0.026)*	*1.021 (0.025)*	*1.035 (0.026)*	*1.075 (0.034)*
ρ_G	*1.025 (0.038)*	*1.042 (0.033)*	*1.025 (0.038)*	*1.027 (0.0039)*
σ^2_ζ	*0.0001 (0.00001)*	*0.0001 (0.00001)*	*0.0001 (0.00001)*	*0.0001 (0.00001)*
σ^2_g	*0.0002 (0.00002)*	*0.0002 (0.00002)*	*0.0002 (0.00002)*	*0.0002 (0.00002)*
		$\chi^2(6) =$ *259.69*	$\chi^2(5) =$*260.19*	$\chi^2(6) =$*257.71*

Table 5.2: Estimates of a RBC model

Given existing experimental evidence and the relatively small sample of data, we decided to estimate θ from a just identified system or from a weakly overidentified one without optimal weighting. Overidentified estimates can be obtained using additional lags of $r_t, \frac{c_{t+1}}{c_t}$, of investment or of the output-labor ratio. Using linearly detrended US quarterly data for the sample 1956:1-1984:1 for consumption, investment, government expenditure, output, household hours and capital stock we estimate θ, fixing φ, which can not be estimated from this data set. Since estimates of η are low, we also provide estimates conditioning on a larger value of η. Table 5.2 reports the results with standard errors in parenthesis (estimates of $\bar{G}$ and $\bar{\zeta}$ are omitted as they are insignificantly different from zero). Four main features emerge from the table. First structural parameters are, in general, precisely estimated and the data wants non-stationary government and technology disturbances. This obtains regardless of whether just-identified or overidentified systems are used and indicates that the model lacks internal propagation. Second, estimates of β are economically unreasonable except when we set $\eta = 0.64$. Third, the model is strongly rejected in all cases with smaller χ^2 statistic when $\varphi = 2$. Fourth, results are broadly independent of the value φ. In fact, apart from ρ_ζ and ρ_G, estimates change very little for φ in the range [0, 3].

While J-test or other statistical devices can give a rough idea of the validity of a model, they are insufficient from an economic point of view since, in the case of failure, they provide no indications on how to respecify the model to improve the fit. Useful information on why the model fails can be obtained by comparing features of the data which have economic content (see chapter 7 for a thorough discussion). Although it is popular to examine these

economic features informally, such comparisons are difficult to interpret since they neglect parameter and sampling uncertainty. Whenever features of interest are continuous functions of the unknown parameters, an (economic) Wald-type test can be employed to formally evaluate the quality of the model's approximation to the data.

Let the features of interest be $\mathsf{S}(\theta)$ and the corresponding features in the data be S_T where the subscript T indicates the sample size. Let $h_T(\theta_T) = \mathsf{S}(\theta_T) - \mathsf{S}_T$ where θ_T is a GMM estimate. Then the covariance matrix of $h_T(\theta_T)$ is $\Sigma_h = (\frac{\partial h(\theta_0)}{\partial \theta'})\Sigma_\theta(\frac{\partial h(\theta_0)}{\partial \theta'})' + \Sigma_{\mathsf{S}}$. Under the null hypothesis that the model reproduces the features of interest in the data $Th(\theta_T)'\Sigma_h^{-1}h(\theta_T) \xrightarrow{D} \chi^2(\dim(\mathsf{S}))$. Hence, a large value of this statistic indicates that the model and the data are different in the dimensions of interest. Since it is possible to run this test for any subset of $\mathsf{S}(\theta)$, the approach can be used sequentially to check which features of the data are matched and which are not.

One important point needs to be emphasized. While statistical tests can be conducted using the optimality conditions of the model, economic tests require a researcher to generate $\mathsf{S}(\theta_T)$ given some estimate θ_T. In other words, to conduct economic tests, a solution to the model needs to be computed. Therefore, one of the main advantages of GMM over maximum likelihood and similar techniques disappears.

Example 5.26 *Continuing with example 5.25, we log linearize the conditions and solve the model. Using just-identified estimates of the parameters we evaluate the quality of the model approximation to the data considering: a) the variance of output, consumption, investment and hours, b) the first order autocorrelation of these four variables, (c) the contemporaneous cross correlations of consumption, investment and hours with output. Table 5.3 reports these statistics for the model and the data. Since* $\mathsf{S} = T \times h(\theta_T)'\Sigma_h^{-1}h(\theta_T) > 800$*, we strongly reject the idea that the RBC model replicates these 11 moments of the data.*

Moment	*Data*	*Model*	*Moment*	*Data*	*Model*	*Moment*	*Data*	*Model*
var(y)	*0.002*	*0.001*	*c-AR(1)*	*0.986*	*0.927*	*corr(c,y)*	*0.953*	*0.853*
var(c)	*0.001*	*0.009*	*inv-AR(1)*	*0.976*	*0.991*	*corr(inv,y)*	*0.911*	*0.703*
var(inv)	*0.005*	*0.008*	*N-AR(1)*	*0.958*	*0.898*	*corr(N,y)*	*0.464*	*0.570*
var(N)	*0.0004*	*0.0003*	*y-AR(1)*	*0.780*	*0.859*			

Table 5.3: Moments of the data and of the model

Exercise 5.43 *(Burnside and Eichenbaum) Let agents' preferences be represented by* $U(c, N, ef) = \ln c_t + \vartheta_N N_{t-1} \ln(1 - b_0 - b_1 ef_t)$ *where* ef_t *is effort,* N_{t-1} *is the probability of working,* ϑ_N *is the fraction of people working,* b_1 *is a parameter and* b_0 *is a fixed costs one has to pay to get to work -* $(1 - b_0 - b_1 ef_t)$ *are effective hours of leisure. Here effort can respond to news instantaneously but* N_t *cannot. We assume a production function of the form* $GDP_t = \zeta_t K_t^{1-\eta}(b_1 ef_t N_{t-1})^\eta$*; that government consumes a random amount* G_t *taxing income at the rate* T^y *and that capital depreciates at the rate* δ*.*
i) Show that in the steady states, if $ef = 1$ *we have* $(GDP/K)^{ss} = \frac{[\beta^{-1} - (1-\delta)]}{1-\eta}$; $(c/GDP)^{ss} =$

$1-\delta(K/GDP)^{ss}-(g/GDP)^{ss}$; $N=\frac{\eta}{\vartheta_N}(1-b_0-b_1)(GDP/c)^{ss}$; $\ln(1-b_0-b_1)=\frac{b_1}{1-b_0-b_1}$.
ii) Show that the first order conditions of the problem can be written as

$$-\vartheta_N b_1 N_{t-1}(1-b_0-b_1 ef_t)^{-1}+\eta c_t^{-1}\frac{y_t}{ef_t} = 0 \quad (5.46)$$

$$\vartheta_N E_t \ln(1-b_0-b_1 ef_{t+1})+E_t c_{t+1}^{-1}\eta\frac{y_{t+1}}{N_t} = 0 \quad (5.47)$$

$$-c_t^{-1}+E_t\beta c_{t+1}^{-1}[(1-\eta)\frac{y_{t+1}}{K_{t+1}}+(1-\delta)] = 0 \quad (5.48)$$

$$\zeta_t K_t^{1-\eta}(b_1 ef_t N_{t-1})^{\eta}+(1-\delta)K_t-K_{t+1}-G_t-c_t = 0 \quad (5.49)$$

iii) Describe how to estimate $(b_0, b_1, \eta, \beta, \delta, \vartheta_N)$ using GMM . Which parameters are identifiable? What data would you use? What instruments would you consider? How would you deal with the fact that effort is non-observable? (Hint: think of a proxy and consider the effects of measurement error).
iv) Test the hypothesis that the model fits the first three terms of the autocovariance function of hours and of the cross covariance of hours and productivity (wage) at lags -1, 0 and 1. Repeat the exercise assuming that effort is fixed (i.e. drop it from the choice variables). Can you test the variable effort model against the fix effort model? How?

Exercise 5.44 *(Eichenbaum and Fisher) Consider monopolistic competitive firms which can't reoptimize their price because information is sticky. This is because, at each t, they only observe variables dated at $t-\tau$.*
(i) Show that log linearizing the optimality conditions leads to $p_{it} = E_{t-\tau}[mc_t+\sum_{l=1}^{\infty}(\beta\zeta_p)^l (mc_{t+l}-mc_{t+l-1}+\pi_{t+l})]$ where β is the discount factor, ζ_p the share of firms not changing prices, mc_t are marginal costs (lower case variables are deviations from the steady state). Show that the (log-linearized) Phillips curve is $\pi_t=\beta E_{t-\tau}\pi_{t+1}+\frac{(1-\beta\zeta_p)(1-\zeta_p)}{\zeta_p}E_{t-\tau}mc_t$.
(ii) Using GDP deflator and (real) labor share data for the US provide GMM estimates of β and ζ_p using as instruments a constant, one/three/five lags of GDP deflator and of the (real) labor share for $\tau=0,1,2$ correcting for serial correlation of order 0 or 2. Provide a test of overidentifying restrictions. Which version of the model fits the data better?
(iii) Repeat (ii) jointly estimating the parameters of the log-linearized Phillips curve and those of the log linearized Euler equation $c_t = E_{t-\tau}[c_{t+1}-\frac{1}{\varphi}(i_{t+1}-\pi_{t+1})]$ where c_t is consumption, i_t the nominal interest rate and φ the coefficient of relative risk aversion. In this case, add lags of consumption and of the nominal interest rates to the instrument list. Do the results in (ii) change?

5.4.1 Some Applied tips

A number of studies have examined the small sample properties of GMM estimators in macro-based or finance-based experimental data (see e.g. Tauchen (1986), Kocherlakota (1990), Mao (1993), Pagan and Yoon (1993) Ferson and Foerster (1994), Hansen et. al. (1996), Newey and West (1996) West and Wilcox (1996), Burnside and Eichenbaum (1996), Den Haan and Christiano (1996), Den Haan and Levin (1996), Anderson and Sörensen

(1996), Furher, Moore and Shuh (1995), Linde (2001) Ruge Murcia (2002)). Four issues have been primarily investigated: (i) how inefficient are estimates obtained using a subset of the available instruments; (ii) how reasonable are two-step GMM estimators in small samples; (iii) how large are the efficiency gains obtained using an optimal weighting matrix; (iv) the relative performance of parametric and kernel based HAC estimates.

On the first issue there is agreement. While a large set of moment conditions improves asymptotic efficiency, it also dramatically increases small sample biases. Hence, GMM estimates obtained with a smaller number of instruments may have lower MSE in small samples. There are two reasons for this: first, additional instruments may only be weakly correlated with the quantities that they instrument for. As we have seen in exercise 5.11, GMM estimators may not even exist if this correlation is weak. Second, when the dimension of the weighting matrix is large, estimates may fail to converge to a non-stochastic matrix in small samples. In general, when the sample is short, one should be careful in taking strongly overidentified estimates at their face value.

For the second issue, the results are mixed and depend on the environment. In general, a fully iterative GMM estimator has good properties for simple problems. However, when the h_t function is highly nonlinear and/or T is relatively small, the small sample distribution may poorly approximate the asymptotic one. Note that in some experimental design, it has been found that fully iterative GMM are poor even when $T = 300$.

As mentioned, estimation of the optimal covariance matrix is complicated, depends on the number of instruments used, the sample size, the serial correlation properties of the h_t function and a number of other choices made by the investigator. In general, estimates of W_T tend to be poor and this, in turn, affects standard error of the estimates and overidentifying tests. When problems are suspected, it may be reasonable to use either the identity matrix or proceed with a just-identified version of the model. Hansen et al. (1996) explored the properties of θ_T obtained minimizing $[\frac{1}{T}\sum_t h(y_t,\theta)]'(W_T(\theta))^{-1}[\frac{1}{T}\sum_t h(y_t,\theta)]$. One reason for preferring a weighting matrix which varies with θ is that under conditional homoskedasticity, θ_T is invariant to how moment conditions are scaled and corresponds to Sargan's IV estimator for a large class of models. It appears that such a choice produces smaller biases in θ_T than a standard selection in some designs, but no other corroborating evidence has been presented in the literature. Also, researchers have found that poor kernel estimates of Σ^+ produce biases, induce small sample confidence intervals with very poor coverage properties and t-tests which overreject the null hypothesis. Den Haan and Levin (1996) have shown that it may be dangerous to entirely rely on automatic bandwidth selection procedures, as they produce outcomes which are hard to believe (e.g that $J(T) = 1$ when $T = 128$). Also, distortions in Σ^+ could also be created because, to insure semi-positiveness, the bandwidth parameter must be the same for each orthogonality condition.

The experimental evidence also suggests that estimates of the parameters and of the standard errors are biased in small samples. The direction of the bias however depends on the design while estimates of the standard errors are, in general, downward biased. Since this implies that t-statistics have long and fat tails, tests of hypotheses when T is small

should be undertaken with caution.

Regarding the reliability of overidentifying tests, the small sample evidence is mixed since results depend on W_T, on whether a fully iterative or a two step GMM is used, on whether instruments carry " good" information, on how many there are, etc. Hence, we recommend experimenting with various alternatives - in particular, with various estimates of W and of instruments - before deriving conclusions about parameter estimates and the quality of the model.

Burnside and Eichenbaum (1996) find that Wald tests in an RBC model overreject individual moment restrictions but that their size increases uniformly as the dimension of the statistics used increases. Also in this case, difficulties are due to poor estimates of the W matrix. Linde' (2001) finds that GMM estimates of New-Keynesian Phillips' curves (obtained from data simulated by a three equation New-Keynesian model) are inaccurate when the forward looking element is strong and when there is measurement error in marginal costs. He also shows that ML estimates are preferable, both in large and small samples.

Two further practical issues are of interest. The asymptotic distribution of GMM estimates was derived under stationarity and ergodicity . We can extend the GMM framework to allow for linear trends in y_t, as in Ogaki (1993), with minor changes. However, the procedure does not allow for unit roots or other forms of nonstationarities in y_t. Hence, it is typical to transform the data (take growth rates) or filter it before estimation is undertaken. In example 5.25 we have eliminated a linear trend, but this did not seem to be enough as estimates of the persistence of the shocks imply processes in the nonstationary region. The alternative is to employ a band pass or a HP filter. As we have seen in chapter 3, filtering is not innocuous. For example, Christiano and Den Haan (1996) find that HP filtering induces large and persistent serial correlation in the residuals of the orthogonality conditions and this creates problems in the estimation of the spectral density at frequency zero. Clearly, problems are more severe when filtered data is very persistent - as is the case with the HP or the band pass filters. If filtering is required, one should compare estimates obtained with different approaches and judgementally select the most reasonable one.

Second, one may wonder how large should T be to be reasonably confident in the results. Experimental evidence on simple specifications suggests that, with T=300, GMM estimators obtained with good W estimates and good instruments approximate the true values in distribution. However, T=300 is a large number: 40 years of time homogeneous quarterly data make T=160 and 25 years of monthly data make T=300. Experimental evidence also suggests that convergence is slow. Hence, caution should be exercised, probably with any available macroeconomic data.

In conclusion, small sample distribution may deviate from the asymptotic one when the weighting matrix is poorly estimated - and this is more of a problem when the orthogonality conditions are highly serially correlated - when the instruments are poorly correlated with the functions we want to instrument for and when too many moment conditions are used in testing relative to the sample size. Hence, when T is small and h_T serially correlated, we recommend a parametric HAC approach or to avoid, when possible, the estimation of W. Also for testing purposes, the number of instruments and/or overidentifying restrictions

should be a function of the sample size.

5.5 Simulation Estimators

Simulation Estimators have become popular over the last 10 years for at least two reasons: they are cheap and easy to compute and they can be used in situations where GMM can't be employed. Two examples where GMM is inapplicable are the following.

Example 5.27 *Suppose in example 5.25 that data on capital is not available. Since equations like (5.38) contain unobservable variables, sample counterparts of theoretical conditions cannot be computed, so GMM can not be employed. One could use the competitive rental rate (approximated by the nominal interest rate minus inflation) in place of $f_K+(1-\delta)$ and still estimate (β,φ) with GMM. However, rejection of the orthogonality conditions is hard to interpret, as it may be due to the approximation employed.*

Example 5.28 *Suppose in example 5.25 that agent's preferences are subject to an unobservable shock υ_t with known distribution. If $u(c_t,N_t,\upsilon_t)=\frac{c_t^{1-\varphi}}{1-\varphi}\upsilon_t+\vartheta_N(1-N_t)$, then*

$$g_\infty(\theta)=E_t[\beta(\frac{c_{t+1}^{-\varphi}\upsilon_{t+1}}{c_t^{-\varphi}\upsilon_t}[f_K+(1-\delta)]-1]=0 \tag{5.50}$$

which is not estimable even if K_t is available, because υ_t is unobservable. However, we can draw $\{\upsilon_t\}^l, l=1,\dots L$ and use $\frac{1}{L}\sum_l[\beta\frac{c_{t+1}^{-\varphi}\upsilon_{t+1}^l}{c_t^{-\varphi}\upsilon_t^l}[f_K+(1-\delta)]-1]=0$ in place of (5.50). In fact, under regularity conditions, $\lim_{L\to\infty}\frac{1}{L}\sum_l[\beta\frac{c_{t+1}^{-\varphi}\upsilon_{t+1}^l}{c_t^{-\varphi}\upsilon_t^l}[f_K+(1-\delta)]-1]\xrightarrow{P}g_\infty(\theta)$.

In general, simulation estimators can be used when h_t contains unobservable variables or shocks. Note that h_t need not be the difference between two orthogonality conditions. In fact, in this section, we let h_t be the difference between generic continuous functions of the parameters in the sample and in the population. Such functions could be orthogonality conditions, moments, VAR coefficients, autocovariances, spectral densities, etc.

5.5.1 The General Problem

Let $x_t(\theta)$ be a $m\times 1$ vector of simulated time series given θ and let y_t be its actual counterpart. Assume that there exists a θ_0 such that $\{x_t(\theta_0)\}_{t=1}^{T_s}$ and $\{y_t\}_{t=1}^{T}$ share the same distribution. Let f be a $n\times 1$ vector of continuous functions; let $F_T(y)=\frac{1}{T}\sum_{t=1}^{T}f(y_t)$ and $F_{T_s}(x,\theta)=\frac{1}{T_s}\sum_{t=1}^{T_s}f(x_t(\theta))$. We would like $F_T(y)\to E[f(y_t)]$ and $F_{T_s}(x,\theta)\to E[f(x_t(\theta))]$ for each θ. If $x_t(\theta)$ and y_t are stationary and ergodic, and f is continuous, convergence obtains almost surely. Furthermore, given the assumptions made, $E[f(y_t)]-f(x_t,\theta_0)]=0$. Given an $n\times n$ random matrix $W_{T,T_s}\xrightarrow{P}W$, $\text{rank}(W_{T,T_s})\geq k$, a simulation estimator θ_{T,T_s} solves:

$$arg\min Q_{T,T_s}=arg\min_\theta[F_T(y)-F_{T_s}(x(\theta))]'W_{T,T_s}[F_T(y)-F_{T_s}(x(\theta))] \tag{5.51}$$

The estimator in (5.51) is similar to the one in (5.2). To show the analogy set

$$h_T(y_t, x_t, \theta) = \frac{1}{T}\sum_{t=1}^{T} f(y_t) - \frac{1}{\kappa}\sum_{t=1}^{T\times\kappa} f(x_t(\theta)) = \frac{1}{T}\sum_{t}^{T}[f(y_t) - \frac{1}{\kappa}\sum_{i=[1+(t-1)\kappa]}^{[\kappa t]} f(x_i(\theta))] \quad (5.52)$$

where $\kappa = \frac{T_s}{T} > 1$ and $[\kappa t]$ is the largest integer less or equal than κt. Then θ_{T,T_s} is a GMM estimator for the h_T function given in (5.52). Note that we can produce a time series for $x_t(\theta)$ of length $T\kappa$, or κ time series all of length T. Which approach one chooses is irrelevant. What is important is that the random numbers used to calculate $x_t(\theta)$ at each replication are fixed since continuity of the objective function may be otherwise violated. Finally, for h_T to be well behaved, we need $\frac{T_s}{T}$ to stay constant as $T, T_s \to \infty$.

Because of the similarities between θ_{T,T_s} and θ_T, the asymptotic properties of θ_{T,T_s} can be obtained by verifying that the general conditions of section 3.1 hold. In particular we need y_t and $x_t(\theta)$ to be mutually independent, stationary and ergodic processes, that h_T has a unique zero, that $f(x_t, \theta)$ and $\partial f(x_t, \theta)/\partial\theta'$ are continuous in the mean at θ_0 and that $F(\theta) = E[\partial f(x_i(\theta))/\partial\theta']$ exists, it is finite and has full rank.

Let $ACF_y(\tau) = E[f(y_t) - E(f(y_t))][f(y_{t-\tau}) - E(f(y_{t-\tau}))]'$, $\Sigma_y = \sum_{-\infty}^{\infty} ACF_y(\tau)$, $ACF_x(\tau) = E[f(x_t(\theta_0)) - E(f(x_t(\theta_0)))][f(x_{t-\tau}(\theta_0)) - E(f(x_{t-\tau}(\theta_0)))]'$; $\Sigma_x = \sum_{-\infty}^{\infty} ACF_x(\tau)$. If the above assumptions are satisfied, we have that

$$\sqrt{T}[F_T(y) - E(f(y_t))] \xrightarrow{D} \mathsf{N}(0, \Sigma_y) \quad (5.53)$$

$$\sqrt{T_s}[F_{T_s}(x(\theta_0)) - E(f(x(\theta_0)))] \xrightarrow{D} \mathsf{N}(0, \Sigma_x) \quad (5.54)$$

and $cov[F_T(y) - F_{T_s}(x(\theta_0))] = \Sigma_y + (1/\kappa)\Sigma_x = (1 + 1/\kappa)\Sigma_y \equiv \overline{\Sigma}$ because $E[f(y_t)] = E[f(x_t(\theta_0))]$. Hence, as $T, T_s \to \infty$, $\frac{T_s}{T}$ fixed, $\sqrt{T}(\theta_{T,T_s} - \theta_0) \xrightarrow{D} \mathsf{N}(0, \Sigma_\theta)$ where $\Sigma_\theta = F(\theta_0)'W(F(\theta_0))^{-1}F(\theta_0)'W\overline{\Sigma}WF(\theta_0)(F(\theta_0)'WF(\theta_0)')^{-1})$.

Exercise 5.45 *i) Show that it is optimal to set $W^\dagger = \overline{\Sigma}^{-1}$. Display the optimal $\Sigma_\theta^\dagger$.*
ii) Show that it is optimal to let $\kappa \to \infty$. (Hint: As $\tau \to \infty$, $\bar{\Sigma} = \Sigma_y$).
iii) Show that a goodness of fit test for model adequacy is $T_1 \times Q_{T,T_s}(\theta_{T,Ts}) \xrightarrow{D} \chi^2(n-k)$.

Exercise 5.46 *Give the form for $\overline{\Sigma}$ when (i) h_T is iid; (ii) h_T is a finite MA process; (iii) h_T is generically serially correlated. Display parametric and non-parametric HAC estimators of $\overline{\Sigma}$ in case (iii). What are the asymptotic properties of these estimators?*

Note that the vector of functions f could be anything a researcher is interested in (e.g. moments, autocorrelations, impulse responses). The only requirement is that f is continuous and that parameters are identifiable. Identifiability, as we will see later, could create some headaches.

Example 5.29 *Suppose that f includes the relative variability of consumption, investment and hours to output in the data and in the model of example 5.25. Then (5.51) defines an*

estimator for at most 3 parameters ($n = 3 \geq k$). By part ii) of exercise 5.45 if the size of simulated time series is sufficiently large (say, $\kappa > 10$), the resulting simulation estimator will be as efficient as GMM - simulation error washes out.

Exercise 5.47 *Consider the setup of exercise 5.2 where $rp_{j,t-1}, j = 1, 2$ are unobservable but known to come from a multivariate normal distribution with mean $\bar{rp}$ and variance Σ_{rp}. Describe how to estimate α_{ij} by simulation. Make sure you specify the function f you employ. What happens to your estimates if expected returns are measured with (iid) errors?*

The setup we have discussed is appropriate in (close to) linear frameworks. In fact, two complications may arise when $x_t(\theta)$ are series generated from a DSGE model. First, draws must be made from the ergodic distribution of x_t which is unknown. Second, the simulated x_t depends on θ in nonlinear way and this implies a nonlinear feedback from parameters to the $f(x_t, \theta)$ function. We illustrate these two issues with an example.

Example 5.30 *(Duffie and Singleton) Let production be $f(K_t, \zeta_t) = \zeta_t K_t^{1-\eta}$ and let the firm maximize the value of dividends by choices of capital i.e., $\max_{\{K_t\}_{t=0}^{\infty}} sd_t = \max_{\{K_t\}_{t=0}^{\infty}} \{\zeta_t K_t^{1-\eta} - r_t K_t\}$ where r_t is the rental rate of capital and ζ_t a technological disturbance. The consumer problem is $\max_{\{c_t, K_{t+1}, S_{t+1}\}_{t=0}^{\infty}} E[\sum_t \beta^t \frac{c_t^{1-\varphi}}{1-\varphi} \upsilon_t]$ subject to $c_t + K_{t+1} + p_t^s S_{t+1} = (sd_t + p_t^s)S_t + (r_t + \delta)K_t$ where υ_t is a taste disturbance, p_t^s is the price of stocks and δ is the depreciation rate. Let $e_t = (\zeta_t, \upsilon_t)'$ be a stationary Markov process with transition function $e_t = P(e_{t-1}, \phi)$ where ϕ is a set of parameters. Let $\theta_0 = (\eta, \beta, \varphi, \delta, \phi)$ and let $y_{2t} = (K_t, e_t)$ be the state vector. In equilibrium y_{2t+1} will be a function of y_{2t}, θ_0 and this mapping may be computed analytically or by simulation. The vector of other endogenous variables y_{1t}, is a function of the states y_{2t}. For example, if $\upsilon_t = 1 \; \forall t$, $\delta = 1$ and $\varphi = 1$, then $K_{t+1} = \beta(1-\eta)\zeta_t K_t^{1-\eta}$; $c_t = (1-\beta(1-\eta))\zeta_t K_t^{1-\eta}$; $sd_t = \eta\zeta_t K_t^{1-\eta}$; $p_t^s = (\beta/(1-\beta)\eta\zeta_t K_t^{1-\eta}$ (see chapter 2).*

Suppose we are not willing to make these assumptions. Then we need to compute y_{2t+1} by simulation, i.e., select a $y_{20} = \bar{y}_2$, $\theta_0 = \bar{\theta}$, draw an iid sequence for the innovations in e_t from some distribution and compute $y_{2i+1}(\bar{\theta})$ recursively. Define $f_t = f(y_{2t}, y_{2t-1}, \ldots, y_{2t-\tau+1})$ and $f_t(\theta) = f(y_{2t}(\bar{\theta}), y_{2t-1}(\bar{\theta}), \ldots, y_{2t-\tau+1})(\bar{\theta})$. Then θ_{T,T_s} minimizes the distance between $F_{T_s}(\theta) = \frac{1}{T_s}\sum_t f_t(\theta)$ and $F_T = \frac{1}{T}\sum_t f_t$. The properties of θ_{T,T_s} are different from the standard simulation estimator framework since $\bar{y}_2$ cannot be drawn from its (stationary) ergodic distribution which is unknown. Hence, the simulated process for y_{2t} depends on the initial conditions and $\bar{\theta}$ and it is therefore nonstationary. If the mapping between y and θ was linear, we could have lessened the problem generating a long time series and throwing away an initial set of observations. However, when the mapping is nonlinear, the dependence on the initial conditions may not die out. Note also that $f_t(\theta)$ depends on θ because of the standard parametric representation and because the transition law of y_{2t} depends on θ. The latter effect is troublesome since $f_t(\theta)$ may not be uniformly continuous.

To take care of these two problems, we impose somewhat stronger conditions on the f function, namely geometric ergodicity and uniform Lipschitz continuity.

Definition 5.3 *(Geometric ergodicity) A time homogeneous Markov process $\{y_t\}_{t=0}^{\infty}$ is geometrically ergodic if for some $b \in (0,1]$, some probability measure μ, (the ergodic distribution of y_t) and for every initial point y_0*

$$b^{-\tau}||P_{t,t+\tau} - \mu||_v \to 0 \quad as \quad \tau \to \infty \tag{5.55}$$

where $P_{t,t+\tau}$ is the τ-step transition probability and $||\Psi||_v \equiv sup_{\{f:|f(y)|\leq 1|\}} \int f(y)d\Psi(y)$ is the total variation norm of the signed measure Ψ.

In words, geometric ergodicity holds if y_t converges at the rate b to the stationary distribution. Note also that geometric ergodicity implies α-mixing with the mixing coefficient converging geometrically to zero. In the discrete case, geometric ergodicity holds if the Markov chain is irreducible and aperiodic and if the mapping between states and parameters and next period states is uniformly convergent. Aperiodicity obtains if the transition matrix does not deterministically alternate between blocks of states (i.e. in a 4×4 matrix with 2×2 blocks we do not have $\begin{bmatrix} 0 & \mu \\ \mu & 0 \end{bmatrix}$). Irreducibility means that each state is accessible from every other state with positive probability. Precise definitions of these two concepts are given in chapter 9.

To dump the effect that θ has on $f_t(\theta)$ via the transition matrix we need the following:

Definition 5.4 *(Uniform Lipschitz condition): A family of functions $\{f_t(\theta)\}$ is Lipschitz, uniformly in probability if there is a sequence $\{b_t\}_{t=1}^{T}$ such that for all $\theta_1, \theta_2 \in \Theta$*

$$||f_t(\theta_1) - f_t(\theta_2)|| \leq b_t||\theta_1 - \theta_2|| \tag{5.56}$$

all t, where $b^T = \frac{1}{T}\sum_{t=1}^{T} b_t$ is bounded in probability.

This condition, together with geometric ergodicity of x_t and a boundness restrictions on the norm of $f_t(\theta)$, implies that the ACF of $f_t(\theta)$ exists and is absolutely summable. In turn, this implies that $f_t(\theta)$ satisfies a weak law of large number and this insures consistency of the simulation estimator for problems like those of example 5.30.

To insure asymptotic normality of the estimator, we need that θ_0 is in the interior of Θ; continuity and differentiability of $f_t(\theta)$, existence and finiteness of $E(\frac{\partial f_t}{\partial \theta'})$. In addition we need that $\Delta_\theta f_t(\theta) = \frac{\partial}{\partial \theta'} f(x_t\theta)$ (the total derivative) satisfies Lipschitz condition uniformly in probability, that $E[|\Delta_\theta f_t(\theta)|] < \infty$ and that $E(\Delta_\theta f_t(\theta))$ is continuous in θ.

Exercise 5.48 *Show that $\sqrt{T}h_T(\theta_0) \xrightarrow{D} \mathsf{N}(0, \bar{\Sigma} = \Sigma_y(1+\kappa^{-1}))$ where h_T was defined in equation (5.52). Show the asymptotic distribution of $\sqrt{T}(\theta_{T,T_s} - \theta_0)$.*

Exercise 5.49 *Show that the asymptotic covariance matrix of the simulation estimator when W is chosen optimally is $\Sigma_\theta^+ = (1+\kappa^{-1})(F_0'\bar{\Sigma}^{-1}F_0)^{-1}$ where $F_0 = E(\frac{\partial f}{\partial \theta'})$. Argue that as $\kappa \to \infty$, Σ_θ^+ approaches the covariance matrix of θ_T and that knowledge of $E(f(\theta))$ increases the efficiency of θ_{T_1,T_2}, unless κ is very large.*

Exercise 5.50 *Suppose f_t is measured with error and let $\tilde{f}_t^{\dagger} = f(x_t, \theta_0) + e_t^f$ where e_t^f is a mean zero, ergodic measurement error. Show that the asymptotic efficiency of a simulation estimator is increased when one ignores the measurement error in simulation.*

Two simulation estimators are popular in the literature. We examine them next.

5.5.2 Simulated Method of Moments Estimator

In a simulated method of moment (SMM) setup one selects θ to minimize the distance between moments of actual and simulated data. Therefore, f_t measures variances, covariances and autocorrelations, etc. In the context of the example 5.25, one could have selected the 11 unknown parameters by simulation using the following algorithm:

Algorithm 5.4

1) *Choose arbitrary values for $\theta = (\beta, \vartheta_N, \varphi, \eta, \delta, \bar{\zeta}, \bar{G}, \rho_\zeta, \rho_g, \sigma_g^2, \sigma_\zeta^2)$ and simulate the model after an (approximate) solution is obtained.*

2) *Let $\mathsf{S} = (\mathsf{S}_1, \mathsf{S}_2)$ be the statistic of interest where S_1 are the conditions dictated by the model - Euler equation, intratemporal conditions, etc.- and S_2 are those selected by the investigator - variances, covariances and autocorrelations. Clearly $dim(\mathsf{S}) \geq 11$ and S_1 could be zero. Compute $\mathsf{S}(\theta) - \mathsf{S}_T$. Update estimates of θ using gradient methods (see Chapter 6).*

3) *Repeat 1)-2) until $||\mathsf{S}(\theta) - \mathsf{S}_T|| < \iota$, ι small.*

SMM is particularly useful when $\mathsf{S}(\theta)$ involves variables with no counterpart in the data. Let $x_t = (x_{1t}, x_{2t})$, let y_{2t} be unobservable with a known distribution. Then, as in example 5.28, one can draw x_{2t} sequences and construct $\mathsf{S}^l(x_{1t}, x_{2t}^l, \theta)$ for each $l = 1, 2, \ldots, L$. If each draw is iid, by the law of large numbers $\frac{1}{L}\sum_{l=1}^{L} \mathsf{S}^l(x_{1t}, x_{2t}^l, \theta) \xrightarrow{P} E_t[\mathsf{S}(x_{1t}, x_{2t}, \theta)]$. Hence, we can use $\mathsf{S}_{lT} = \frac{1}{L}\sum_l \mathsf{S}^l(x_{1t}, x_{2t}^l, \theta)$ in place of the unknown $\mathsf{S}(\theta)$ function so long as L is large enough.

Exercise 5.51 *Consider a log-linearized version of the model of example 5.25 and suppose you choose parameters to match the cross-covariance function of hours and productivity. Since three parameters can be obtained from the moments of the government expenditure process, you have 8 free parameters. Select three autocovariance of each of the two series and three cross covariances. Using the same data as in example 5.25 provide SMM estimates of the free parameters and a test for overidentification.*

Example 5.31 *We reconsider the New-Keynesian Phillips curve of example 5.11. Assuming $\kappa = 1$, we estimate $\theta = (\beta, \zeta_p)$ so as to make the variance and the first two autocorrelations of inflation in the data and in the model as close as possible. Using CPI inflation and the output gap in the three countries, a grid of 100 values for $\beta = [0.98, 1.02]$*

and $\zeta_p = [0.20, 0.98]$ *we obtain* $\beta = (0.986, 1.009, 1.011)$*, and of* $\zeta_p = (0.155, 0.115, 0.422)$ *for US, UK and Germany, respectively. The values of the criterion function at* θ_{T_1,T_2} *are* $(23.32, 114.14, 37.89)$ *indicating, perhaps unsurprisingly, that the model fails to replicate the variability and the AR structure of actual inflation in these countries.*

Exercise 5.52 *Consider the setup of exercise 5.43 but assume that effort is unobservable. Provide an algorithm to obtain SMM estimates of the free parameters. Which moments would you consider? Which instruments? Are there parameters which are not identifiable?*

5.5.3 Simulated Quasi-Maximum Likelihood/ Indirect Inference

The method of simulated quasi-maximum likelihood (SQML) is useful when a researcher is interested in approximating the conditional density of the data. Since a VAR with iid errors can capture well this density if T is large, SQML can be thought as selecting θ so as to match the VAR representation of actual and simulated data.

Let the conditional density of simulated data be $f(x_t(\theta)|x_{t-1}(\theta), \ldots, x_{t-q}(\theta), \alpha)$, where $\alpha \in R^{k'}$ are "shallow" parameters, $k' \geq k$. Note that f may be misspecified in the sense that the true conditional density of $x_t(\alpha)$ may not belong to the set of functions $\{f(x_t(\theta)|x_{t-1}(\theta), \ldots, x_{t-p}(\theta), \alpha)\}$. In principle, one could choose f to approximate as best as possible the true conditional density but in practice computational considerations suggest to select f so that it is easy to obtain a quasi-maximum likelihood estimate of α. When f is a VAR with iid errors, α includes the VAR coefficients on lagged variables and the parameters of the covariance matrix. Hence, while the structural model may be highly non-linear in θ, the estimated model for $x_t(\theta)$ is linear in α.

The quasi-log likelihood of the model is $\mathcal{L}_{T_s}(\{x_t(\theta)\}, \alpha) \equiv \sum_{t=1}^{T_s} logf(x_t(\theta), \ldots, x_{t-q}(\theta), \alpha)$. We let $\alpha_{T_s}(\theta) \equiv argmax\mathcal{L}_{T_s}(\{x_t(\theta)\}, \alpha)$. Since there is no closed form expression, the mapping between θ and α needs to be computed by simulation. If T_s is sufficiently large, $\alpha_{T_s}(\theta) \xrightarrow{p} \alpha(\theta)$. Let $\mathcal{L}_T(\{y_t\}, \alpha) \equiv \sum_{t=1}^{T} logf(y_t, \ldots, y_{t-p}, \alpha)$ be the quasi-log likelihood for the actual data and let $\alpha_T \equiv argmax\mathcal{L}_T(\{y_t\}, \alpha)$. If T is sufficiently large, $\alpha_T \xrightarrow{p} \alpha$. We set $T_s = \kappa T,\ \kappa \geq 1$.

We assume that there exists a θ_0 such that $\alpha_0 = \alpha(\theta_0)$ (this condition is typically referred as encompassing). This does not mean the theoretical model is a good representation of the data: instead it simply requires the much weaker condition that there is a set of structural parameters which makes the "shallow" parameters computed from actual and simulated data identical. Then, a simulated quasi-maximum likelihood (SQML) estimator θ_{T,T_s} of θ_0 solves:

$$\theta_{T,T_s} \equiv argmax_\theta \mathcal{L}_T(\{y_t\}, \alpha_{T_s}(\theta)) \tag{5.57}$$

In words, we maximize the likelihood function of the actual data once we plug in the shallow parameters obtained from maximizing the likelihood function of the simulated data.

Example 5.32 *(Consumption function) Suppose a researcher is interested in finding parameters so that the consumption function generated by a RBC model matches the one in the data. Let the actual data be represented by a bivariate normal VAR(1) in consumption*

and output. Let α include the four VAR coefficients and the three coefficients of the covariance matrix of the errors and let θ include all the parameters of the model. Then the following algorithm can be used:

Algorithm 5.5

1) *Choose a θ_1 and simulate $x_t(\theta_1)$, $t = 1, \ldots, T_s$.*

2) *Fit a VAR(1) to simulated consumption and output and obtain $\alpha_{T_s}(\theta_1)$.*

3) *Use $\alpha_{T_s}(\theta_1)$ in the quasi-log likelihood function of y_t, i.e. compute VAR residuals using the actual data and $\alpha_{T_s}(\theta_1)$ and construct the log likelihood using the prediction error decomposition (see chapter 7).*

4) *Update θ_1 and repeat step 1)-3) until $||\mathcal{L}_T(\{y_t\}, \alpha_{T_s}(\theta_1^l)) - \mathcal{L}_T(\{y_t\}, \alpha_{T_s}(\theta_1^{l-1}))|| \leq \iota$, or $||\theta_1^l - \theta_1^{l-1}|| < \iota$ ι small.*

Note that if $k' \geq k$, the SQML estimator maximizes the quasi-log likelihood function subject to a set of $k' - k$ (nonlinear) restrictions. If the inequality is strict, there are $k' - k$ overidentifying restrictions which can be used to test the quality of the model. A bivariate VAR(3) without a constant, has, e.g., $(2 \times 3) * 2 + 3$ parameters. If $dim(\theta) = 5$, there are 10 testable restrictions.

At times the distinction between SQML and SMM is blurred as the next example shows.

Example 5.33 *In exercise 5.51 estimates are obtained matching the cross covariance function of hours and productivity. If we represent actual and simulated data with a bivariate VAR, we can compute the cross covariance function using the companion form. That is, if $\mathsf{Y}_t = \mathsf{A}\mathsf{Y}_{t-1} + \mathsf{E}_t$, $var(\mathsf{Y}_t) = (I - \mathsf{A})^{-1}\Sigma_E((I - \mathsf{A})^{-1})'$ and $cov(\mathsf{Y}_t, \mathsf{Y}_{t-\tau}) = \mathsf{A}^\tau var(\mathsf{Y}_t)$.*

When the "shallow" parameters are not the coefficients of the VAR representation of the data, SQML is typically termed indirect inference principle. Here it is typical to split $\theta = (\theta_1, \theta_2)$, where θ_2 are nuisance parameters needed for simulations but uninteresting from an economic viewpoint. Let f_T and $f_{T_s}(\theta)$ be vectors of shallow functions in actual and simulated data. Dridi and Renault (1998) showed the following two results:

Result 5.1 *If there exists a $\bar{\theta}_2 \in \Theta_2$ such that $\lim_{T,T_s \to \infty} f_T^0 = f_{T_s}(x_t, \theta_1^0, \bar{\theta}_2)$, and $W_T \xrightarrow{p} W$ $\theta_{1,T,T_s} = arg\min(f_T - f_{T_s}(x_t, \theta_1, \bar{\theta}_2))'W_T(f_T - f_{T_s}(x_t, \theta_1, \bar{\theta}_2))$ is consistent for θ_1^0.*

Result 5.2 $\sqrt{T}(\theta_{T,T_s} - (\theta_1^0, \bar{\theta}_2)') \xrightarrow{D} \mathsf{N}(0, (\frac{\partial[f_T - f_{T_s}(x_t, \theta_1, \bar{\theta}_2)]}{\partial \theta_1'})'\Sigma_0^{-1}\frac{\partial[f_T - f_{T_s}(x_t, \theta_1, \bar{\theta}_2)]}{\partial \theta'})^{-1})$; *where Σ_0 depends on $var(f_T)$ $var(f_{T_s})$, $cov(f_T, f_{T_s})$, $cov(\theta_{T,T_s}^l, \theta_{T,T_s}^{l'})$ and l, l' refer to simulations.*

Note that if there are no nuisance parameters, the condition of result 5.1 collapses to a standard encompassing one $E(f_T - f_{T_s}(x_t, \theta^0)) = 0$. Also, the conditions of results 5.1 and 5.2 are only sufficient. For necessary conditions, see e.g. Gourieroux and Monfort (1995).

Example 5.34 *(Merha and Prescott) Consider the equity premium puzzle popularized by Merha and Prescott (1985). Here* $f_T = (\frac{1}{T}\sum_t R_t^f, \frac{1}{T}\sum_t EP_t)$*, the average risk free rate and the average equity premium;* θ_2 *represents the mean, the variance and the persistence of the endowment process; while* $\theta_1 = (\beta, \varphi)$ *are the parameters of preferences. The literature has tried to (informally) find the range of* θ_1 *such that* $f_{T_s}(\theta_1, \bar{\theta}_2)$ *are as close as possible to* f_T*, given some estimate of* θ_2*. A puzzle is generated because when* $T_s = T$ *values for* θ_1 *in a reasonable range produce* $|f_T - f_{T_s}(x_t, \theta_1, \bar{\theta}_2)|$ *which is large.*

Example 5.35 *(Canova and Marrinan) It is typical to find that the forward rate (fer) is a biased predictor of the future spot exchange rate (ner). That is, in the regression* $ner_{t+1} = \alpha_0 + \alpha_1 fer_{t,1} + e_{t+1}$ *estimates of* α_0 *and* α_1 *significantly differ from 0 and 1, respectively. One question of interest is whether this bias is consistent with optimizing agents and rational expectations. Suppose we can simulate* $ner^s_{t+1}(\theta)$ *and* $fer^s_{t,1}(\theta)$ *from a model and run the regression* $ner^s_{t+1}(\theta) = a_0 + a_1 fer^s_{t,1}(\theta) + e^s_{t+1}$*. Then we can ask if there is a range of* θ *such that* $\alpha_0 = a_0$ *and* $\alpha_1 = a_1$ *or, at least, such that the sign of* (α_0, α_1) *is the same as that of* (α_0, α_1)*.*

Exercise 5.53 *One theory of the term structure of interest rates suggests that the return obtained on a long term bond is a weighted average of the returns on successive short term ones. Using a version of the model considered in example 5.25 obtain indirect inference estimates of the structural parameters and of the parameters of the technology process so that the coefficients in the regressions* $R_{t,t+4} = \alpha_1 + \alpha_2 R_{t,1} + \alpha_3 R_{t+1,1} + \alpha_4 R_{t+2,2} + e_{1t}$ *and* $R_{t,t+2} = \alpha_5 + \alpha_6 R_{t,1} + \alpha_7 R_{t+1,1} + e_{2t}$ *obtained in the model and in the data are the same where* $R_{t,t+4}$ *are returns on one year bonds and* $R_{t,t+1}(R_{t,t+2})$ *are returns on a 90 and 180 days T-bills. (Hint: you need to impose more conditions to estimate all the parameters: several are not identifiable from these regressions. Also, do not use a log-linear approximation to solve this problem). Can the model match the short end of term structure of interest rates?*

Example 5.36 *We use an indirect inference estimator to estimate the parameters of the New-keynesian Phillips curve of example 5.11 using US data. Here the functions we match are the regression coefficients in* $\pi_{t+1} = \alpha_1 \pi_t - \alpha_2(gap_t) + e_{t+1}$*. We present results in table 5.4 for two specifications: one where we use the actual output gap in the simulation and one where a process for the output gap is estimated using an AR(2) and a constant on HP filtered data and then simulated. Standard errors are in parenthesis. The model can roughly replicate the magnitude of* α_1 *found in actual regression. Note that, because* α_2 *is poorly estimated, we have hard time to produce the correct sign for this coefficient when the actual gap is used. Note also that estimated* ζ_p *are very low (roughly, prices change every 1-2 quarters) and that* β *is unreasonably low when the actual gap is used.*

Exercise 5.54 *(Bayraktar, Sakellaris, Vermeulen) Consider the investment decision of a monopolistic competitive firm. Output is produced* $\zeta_{it} K_{it}^{1-\eta}$ *where* ζ_{it} *a technology shock which includes both individual and aggregate components. Suppose the firm chooses capital and borrowing to maximize profits and suppose there are convex costs* $\frac{b_1}{2}(\frac{inv_t}{K_t})^2 K_t$ *and fixed*

	α_1	α_2	β	ζ_p	*Criterion function*
Actual	*0.993 (0.05)*	*-0.04 (0.143)*			
Simulated (actual gap)	*0.996(0.0062)*	*0.032 (0.001)*	*0.752*	*0.481*	*0.01012*
Simulated (simulated gap)	*0.997(0.00008)*	*-0.004 (0.0006)*	*0.980*	*0.324*	*0.02321*

Table 5.4: Indirect Inference Estimates of New Keynesian Phillips curve

costs b_2K_t to adjust capital. Suppose that investment is partially reversible so that the selling price of capital (p^{ks}) is lower than the buying price of capital (p^{kb}) and suppose there exists an external finance premium of the form $b_3\frac{B_t}{p^{ks}K_t}$ where B_t are borrowing and b_1, b_2, b_3 are parameters (if $B_t < 0, b_3 = 0$). The choice of the firm is partially discrete (it must select an action between buying capital, selling capital or doing nothing) and partially continuous (select B_{t+1}). The value function assocaited with each choice is $\mathsf{V}^j(\zeta, K, B) = \max_{\{K^+,B^+\}}(\zeta K^{1-\eta} - C^j(K, inv) + B^+ - (1+r)(1+b_3\frac{B}{p^{ks}K}B) + \beta E\mathsf{V}^*(\zeta^+, K^+, B^+$ where

$$C^j(K_t, B_t) = p^{kb}inv_t + \frac{b_1}{2}(\frac{inv_t}{K_t})^2K_t + b_2K_t \quad \text{if } inv_t > 0 \tag{5.58}$$

$$= p^{ks}inv_t + \frac{b_1}{2}(\frac{inv_t}{K_t})^2K_t + b_2K_t \quad \text{if } inv_t < 0 \tag{5.59}$$

$$= 0 \quad \text{if } inv_t = 0 \tag{5.60}$$

subject to $inv = K^+ - (1-\delta)K$, where "+" indicates future values. The structural parameters are $\theta = (\beta, \delta, \eta, b_1, b_2, b_3, p^{ks}, p^{kb})$ and $(\rho_\zeta, \sigma^2_\zeta)$. Using quarterly data for aggregate output, investment, capital and total bank borrowing and the regression $(inv_t - \bar{inv}) = \alpha_0, +\alpha_1(GDP_t - G\bar{D}P) + \alpha_2(GDP_t - G\bar{D}P)^2 + \alpha_3\frac{B_t-\bar{B}}{K_t-\bar{K}} + \alpha_4\frac{(\zeta_t-\bar{\zeta})(B_t-\bar{B}_t)^2}{K_t-\bar{K}} + e_{it}$, where bar variables are time averages, find indirect inference estimates of $(b_1, b_2, b_3, p^{ks}, p^{kb})$ assuming $r_t = r = 0.02, \beta = 0.99, \delta = 0.025, \eta = 0.66$. Compute moments and compare them to those obtained using optimal parameters.

Exercise 5.55 *(Martin and Pagan) A two state Markov Switching model for y_t can be written as $y_t = \theta_0 + y_{1t}$, $y_{1t} = (\theta_1 + \theta_2 y_{2t})^{0.5}e_{1t}$ and $y_{2t} = (1-p_2) + (p_1 + p_2 - 1)y_{2t-1} + [(p_2(1-p_2)) + (p_1(1-p_1) - p_2(1-p_2))y_{2t-1}]^{0.5}e_{2t}$ where $e_{1t} \sim \mathsf{N}(0,1)$, e_{2t} can take two values and $p_2 = P[y_{2t} = 0|y_{2t-1} = 0]$, $p_1 = P[y_{2t} = 1|y_{2t-1} = 1]$. Suppose that p_1, p_2 are known. Consider a "shallow" function $\alpha = \alpha(\theta_0, \theta_1, \theta_2)$ and suppose α is obtained by solving $E(\sum_t \frac{\partial \mathcal{L}}{\partial \alpha'}) = 0$ where $\mathcal{L}$ is the likelihood function of an auxiliary model involving y_t and α. Using data for US output, obtain indirect inference estimates of a_0, a_1, a_2.*

Example 5.37 *It is common in to derive optimal monetary policy rules. Such rules typically involve full commitment or some kind of cooperative device which are not implementable in competitive economies. Hence researchers have approximated the optimal response of the endogenous variables of the model with simple policy rules which involve feedback from observables to nominal interest rates. Let f_T be a set of optimal impulse responses*

of the model and let $f_{T_s}(y,\theta)$ be the set of simulated responses where θ represents the parameters of the Taylor rule. A $\theta_{T_1T_2}$ which minimizes the distance between optimal and suboptimal (but implementable) responses to some shock is an indirect inference estimator.

Exercise 5.56 *Consider matching output and price responses to monetary policy shocks identified in the data by the requirement that when interest rate increases, real balances decline. Using the sticky price model of example 2.18 of chapter 2, find indirect inference estimates of the parameters that come as close as possible to match the first 10 responses of prices and output obtained in the data (Hint: Make sure that the parameters of the money demand function imply a negative correlation between real balances and interest rates in response to monetary shocks, and choose the remaining parameters to match responses).*

The Indirect inference principle naturally links to calibration (see chapter 7). To make the link explicit we need to define a partial Indirect Inference Estimator. Such an estimator is obtained if only some of the components of $f_{T_s}(x_t,\theta)$ (among the many that the model provides) are used for estimation. That is, we assume that there exists a θ_1^0 such that $f_T^1 = f_{T_s}^1(\theta_1^0,\bar{\theta}_2)$ with $f_T^1 \subset f_T$. This estimator is semi-parametric, since not all the features of the model are fully specified, and $f_T \neq f_{T_s}(\theta_0^1,\bar{\theta}_2)$ that is, the model is potentially misspecified in some dimensions. Then θ_{1,T,T_s} minimizes $Q_{T,T_s}^1(\bar{\theta}_2) = [f_T^1 - f_{T_s}^1(x_t,\theta_1,\bar{\theta}_2)]' W_{(1,T_1,T_2)}[f_T^1 - f_{T_s}^1(x_t,\theta_1,\bar{\theta}_2)]$ and as $T \to \infty$, $T \times Q_{T,T_s}^1(\bar{\theta}_2) \xrightarrow{D} \chi^2(dim(f_T^1) - dim(\theta_1))$. This asymptotic distribution is valid if $\bar{\theta}_2$ is replaced by a θ_{2,T,T_s} satisfying $\sqrt{T}(\theta_{2,T,T_s} - \bar{\theta}_2) \xrightarrow{P} 0$.

Chapter 6: Likelihood methods

Maximum likelihood (ML) techniques have enjoyed a remarkable come back in the last few years, probably as a consequence of the development of faster computer technology and of the substantial improvement in the specification of structural models. In fact, complex stochastic general equilibrium models have been recently estimated and tested against the data. This represents a shift of attitude relative to the 1980's or the beginning of the 1990's where GMM and related techniques dominated the scene. As we have seen maximum likelihood is a special case of GMM when the scores of the likelihood are used as orthogonality conditions. Nevertheless, (full information) ML differs from GMM in several respects.

In both cases, a researcher starts from a fully specified dynamic stochastic general equilibrium model. However, while with GMM the first order conditions of the maximization are sufficient for estimation and testing, with maximum likelihood the final form, expressing the endogenous variables of the model as a function of the exogenous variables and of the parameters, is needed. As we have seen in Chapter 2, this is not a small enterprise in general and approximations are often needed, transforming nonlinear specifications into a linear ones. The presence of nonlinearities, on the other hand, does not present particular problems for GMM estimation and testing. Moreover, while with GMM one uses only the (limited) information contained in a subset of the equilibrium conditions, e.g. the Euler equations, once the final form is calculated, all the implications of the model must necessarily be taken into account for estimation. Therefore, while with the former one can estimate and test assuming that only some of the equations of the model appropriately characterize the data generating process, such an assumption is untenable when ML is used. An interesting conundrum arises when misspecification is present. Following White (1982), one can show that a quasi-ML estimator of the parameters, obtained when the distribution of the errors is misspecified, has the same asymptotic properties as the correct ML estimator under a set of regularity conditions. However, as we will argue in chapter 7, the misspecification present in DSGE models is unlikely to be reducible to the distributions of the errors. Hence, it is unknown what kind of properties ML estimates have in these setups and care must be used in reporting and interpreting estimates and tests.

With both ML and GMM the final scope of the analysis is the evaluation of the quality of the model's approximation to the data and, given estimates, to study the effects of altering interesting economic (policy) parameters. This should be contrasted with the exercises typically performed in VARs. Here the full implications of the model, as opposed to a set of

minimal restrictions, are used to obtain estimates of the objects of interest; the analysis is geared towards the estimation of "structural parameters' as opposed to "structural shocks"; and model evaluation is often more important than describing the (restricted) structure of the data in response to disturbances. Which approach one subscribes depends on how much a researcher trusts the model. With ML (and GMM) one puts a lot of faith in the model as a description of the data - the structure is correct, only the parameters are unknown. With VARs the opposite is true. Therefore only a limited set of conventional or generic restrictions are considered.

This chapter describes the steps needed to estimate models with ML. We start by describing the use of the Kalman filter and of the Kalman smoother for state space models. State space models are general structures: any multivariate ARMA model and almost all log-linearized DSGE model can be fit into this framework. The Kalman filter, besides providing minimum MSE forecasts of the endogenous variables and optimal recursive estimates of the unobserved states, is an important building block in the prediction error decomposition of the likelihood. In fact, the likelihood function of a state space model can be conveniently expressed in terms of the one-step ahead forecast errors, conditional on the initial observations, and of their recursive variance, both of which can be obtained with the Kalman filter. Therefore, given some initial parameter values, the Kalman filter can be used to recursively construct the likelihood function; gradient methods can be employed to provide new estimates for the parameters and the two-step procedure can be repeated until the gradient or the parameters do not change across iterations.

In the third section we provide some numerical tips on how to update parameter estimates and on other issues often encountered in practice. The algorithms are only sketched here. For details the reader should consult Press et al. (1980) or Judge, et. al (1985). The last portion of this chapter applies the machinery we have developed to the problem of estimating DSGE models. The (log)-linearized solution of such models naturally comes into a state space format where the coefficients are highly nonlinear functions of the structural parameters. We discuss a number of peculiarities of DSGE models relative to other time series specifications and describe how to use cross-equations restrictions to identify structural parameters and to test the model. This is the approach popularized by Sargent (1979) and Sargent and Hansen (1980) and exploits the fact that linearized expectational equations impose restrictions on the VAR of the data. We conclude estimating the parameters of a simple sticky price model driven by technology and monetary disturbances and confronting some of the implied unconditional moments to the data.

6.1 The Kalman filter

The Kalman filter is one of the most important instruments in the toolkit of applied macroeconomists and we will extensively use it throughout the rest of this book. The presentation here is basic and the reader should refer to Harvey (1991) or Anderson and Moore (1979) for more extensive details.

The Kalman filter is typically employed in state space models of the form

$$y_t = x'_{1t}\alpha_t + x'_{2t}v_{1t} \tag{6.1}$$
$$\alpha_t = \mathbb{D}_{0t} + \mathbb{D}_{1t}\alpha_{t-1} + \mathbb{D}_{2t}v_{2t} \tag{6.2}$$

where x'_{1t} is $m \times m_1$ matrix, x'_{2t} is $m \times m_2$ matrix , $\mathbb{D}_{0t}$ is $m_1 \times 1$ vector, $\mathbb{D}_{1t}$, $\mathbb{D}_{2t}$ are $m_1 \times m_1$ and $m_1 \times m_3$ matrices; v_{1t} is a $m_2 \times 1$ vector of martingale difference sequences, $v_{1t} \sim \mathbb{N}(0, \Sigma_{v_1})$; v_{2t} is $m_3 \times 1$ vector of martingale difference sequences, $v_{2t} \sim \mathbb{N}(0, \Sigma_{v_2})$. We also assume that $E(v_{1t}v'_{2\tau}) = 0$ and $E(v_{1t}\alpha'_0) = 0$, for all t and τ. The first assumption can be dispensed of, as we will see later on. The two together insure that the states α_t and the disturbances v_{1t} are uncorrelated.

(6.1) is typically referred as the measurement (observation) equation while (6.2) is the transition (state) equation. Note that, in principle, α_t is allowed to vary over time and that $x_{1t}, x_{2t}, \mathbb{D}_{0t}, \mathbb{D}_{1t}, \mathbb{D}_{2t}$ could be fixed (i.e. matrices of numbers) or realizations of random variables. For example, in time series context x_{1t} could contain lagged y_t's and x_{2t} current and/or lagged stochastic volatility terms. Notice that it is possible to have m_2 shocks driving the m endogenous variables, $m_2 \leq m$.

The framework provided by (6.1)-(6.2) is general: a number of time series and regression models can be cast in such a format. We consider a few special cases next.

Example 6.1 *Consider an m variable VAR $y_t = A(\ell)y_{t-1} + e_t$, where $A(\ell)$ is a polynomial of order q and e_t is a martingale difference process, $e_t \sim (0, \Sigma_e)$. As we have seen such a system can be rewritten in a companion form as $\mathbb{Y}_t = \mathbb{A}\mathbb{Y}_{t-1} + E_t$ where $\mathbb{A} = [\mathbb{A}_1, \mathbb{A}_2]'$ and $\mathbb{A}_1 = (A_1, \ldots, A_q)'$ contains the first m rows of $\mathbb{A}$, $\mathbb{A}_2$ is a matrix of ones and zeros and $E_t = (e_t, 0, \ldots, 0)'$. Such a system fits into (6.1)-(6.2) setting $\alpha_t = \mathbb{Y}_t = [y'_t, y'_{t-1}, \ldots y'_{t-q}]'$, $x'_{1t} = [I, 0, \ldots 0]$, $\mathbb{D}_{1t} = \mathbb{A}, \Sigma_{v_1} = 0$, $v_{2t} = E_t$, $\mathbb{D}_{2t} = I$, $\mathbb{D}_{0t} = 0$. Hence, there is no measurement error, the measurement equation is trivial and states and observables coincide.*

Example 6.2 *Consider the univariate process, $y_t = A_1 y_{t-1} + A_2 y_{t-2} + e_t + D_1 e_{t-1}$. This model can be equivalently written as:*

$$y_t = [1\ 0] \begin{bmatrix} y_t \\ A_2 y_{t-1} + D_1 e_t \end{bmatrix}$$

$$\begin{bmatrix} y_t \\ A_2 y_{t-1} + D_1 e_t \end{bmatrix} = \begin{bmatrix} A_1 & 1 \\ A_2 & 0 \end{bmatrix} \begin{bmatrix} y_{t-1} \\ A_2 y_{t-2} + D_1 e_{t-1} \end{bmatrix} + \begin{bmatrix} 1 \\ D_1 \end{bmatrix} e_t$$

Hence, an ARMA(2,1) structure fits (6.1)-(6.2) setting $\alpha_t = \begin{bmatrix} y_t \\ A_2 y_{t-1} + D_1 e_t \end{bmatrix}$, $\mathbb{D}_{1t} = \begin{bmatrix} A_1 & 1 \\ A_2 & 0 \end{bmatrix}$, $\mathbb{D}_{2t} = \begin{bmatrix} 1 \\ D_1 \end{bmatrix}$, $\mathbb{D}_{0t} = 0$, $x'_{1t} = [1,\ 0], \Sigma_{v_1} = 0, \Sigma_{v_2} = \sigma_e^2$.

Exercise 6.1 *Consider a process of the form $y_{1t} = A_1(\ell)y_{1t-1} + D(\ell)e_t + A_2 y_{2t}$ where y_{2t} represents exogenous variables, $A_1(\ell)$ is of order q_1 and $D(\ell)$ of order q_2. Show the form of the state space model in this case. Display $\mathbb{D}_{1t}, \mathbb{D}_{2t}, x'_{1t}, x'_{2t}$.*

Besides time series models, several structures naturally fit into a state space framework.

Example 6.3 *1) In many economic problems the ex-ante real rate is needed but only the ex-post real rate of interest is computable. In this case we could set* $\alpha_t \equiv r_t^e = i_t - \pi_t^e$ *where* π_t^e *is the expected inflation rate and assume e.g.,* $\alpha_t = \mathbb{D}_1\alpha_{t-1} + v_{2t}$*. The observed real rate then is* $y_t \equiv i_t - \pi_t = \alpha_t + v_{1t}$ *where* v_{1t} *is a measurement error.*
2) A RBC model driven by unit root technology shocks implies that all endogenous variables have a common trend (see King, Plosser, Stock and Watson (1991)). Here $\alpha_t = \alpha_{t-1} + v_{2t}$ *is a one dimensional process;* $x'_{1t} = x'_1$ *are the loadings on the trend and* $x'_{2t} = x'_2$ *are the loadings on everything else (cycle, irregular, etc.).*

Exercise 6.2 *When agents are risk neutral, uncovered interest parity implies that interest rates differentials should be related to the expected change in the exchange rate (see example 2.3 of chapter 5). Cast such a relationship into a state space format carefully defining the matrices* $x'_{1t}, x'_{2t}, \mathbb{D}_{0t}, \mathbb{D}_{1t}, \mathbb{D}_{2t}$*.*

Exercise 6.3 *(Nonlinear state space model) Consider the model* $y_t = \alpha_t + v_{1t},\ \alpha_{t+1} = \alpha_t\theta + v_{2t}$ *and suppose one is interested in* θ*, which is unobservable, as is* α_t*. (In a trend-cycle decomposition,* θ *represents, e.g., the persistence of the trend). Cast the problem in a state space format; show the state vector and display the matrices of the model.*

The Kalman filter can be used to optimally estimate the unobservable state vector α_t and to update estimates when a new observation becomes available. As a byproduct, it also produces recursive forecasts of y_t, consistent with the information available at t.

Suppose we want to compute $\alpha_{t|t}$, the optimal (MSE) estimator of α_t using information up to t; and $\Omega_{t|t}$ the MSE matrix of the forecast errors in the state equation. At this stage we let $x'_{1t} = x'_1,\ x'_{2t} = x'_2, \mathbb{D}_{1t} = \mathbb{D}_1,\ \mathbb{D}_{0t} = \mathbb{D}_0,\ \mathbb{D}_{2t} = \mathbb{D}_2$ be known. We also assume that a sample $\{y_t\}_{t=1}^T$ is available. The Kalman filter algorithm has five steps.

Algorithm 6.1

1) *Select initial conditions. If all eigenvalues of* $\mathbb{D}_1$ *are less then one in absolute value, set* $\alpha_{1|0} = E(\alpha_1)$ *and* $\Omega_{1|0} = \mathbb{D}_1\Omega_{1|0}\mathbb{D}'_1 + \mathbb{D}_2\Sigma_{v_2}\mathbb{D}'_2$ *or* $vec(\Omega_{1|0}) = (I - (\mathbb{D}_1 \otimes \mathbb{D}'_1)^{-1})vec(\mathbb{D}_2\Sigma_{v_2}\mathbb{D}'_2)$*, in which case the initial conditions are the unconditional mean and variance of the process. When some of the eigenvalues of* $\mathbb{D}_1$ *are greater than one, initial conditions cannot be drawn from the unconditional distribution and one needs a guess (say,* $\alpha_{1|0} = 0$*,* $\Omega_{1|0} = \kappa * I$*,* κ *large) to start the iterations.*

2) *Predict* y_t *and construct the mean square of the forecasts using* $t-1$ *information*

$$\begin{aligned} E(y_{t|t-1}) &= x'_1\alpha_{t|t-1} \qquad (6.3) \\ E(y_t - y_{t|t-1})(y_t - y_{t|t-1})' &= E(x'_1(\alpha_t - \alpha_{t|t-1})(\alpha_t - \alpha_{t|t-1})'x_1) + x'_2\Sigma_{v_1}x_2 \\ &= x'_1\Omega_{t|t-1}x_1 + x'_2\Sigma_{v_1}x_2 \equiv \Sigma_{t|t-1} \qquad (6.4) \end{aligned}$$

3) *Update state equation estimates (after observing y_t):*

$$\alpha_{t|t} = \alpha_{t|t-1} + \Omega_{t|t-1}x_1\Sigma_{t|t-1}^{-1}(y_t - x_1'\alpha_{t|t-1}) \quad (6.5)$$

$$\Omega_{t|t} = \Omega_{t|t-1} - \Omega_{t|t-1}x_1\Sigma_{t|t-1}^{-1}x_1'\Omega_{t|t-1} \quad (6.6)$$

where $\Sigma_{t|t-1}^{-1}$ is defined in (6.4).

4) *Predict the state equation random variables next period:*

$$\alpha_{t+1|t} = \mathbb{D}_1\alpha_{t|t} + \mathbb{D}_0 = \mathbb{D}_1\alpha_{t|t-1} + \mathbb{D}_0 + \mathsf{K}_t\epsilon_t \quad (6.7)$$

$$\Omega_{t+1|t} = \mathbb{D}_1\Omega_{t|t}\mathbb{D}_1' + \mathbb{D}_2\Sigma_{v_2}\mathbb{D}_2' \quad (6.8)$$

where $\epsilon_t = y_t - x_1'\alpha_{t|t-1}$ is the one-step ahead forecast error in predicting y_t, and $\mathsf{K}_t = \mathbb{D}_1\Omega_{t|t-1}x_1\Sigma_{t|t-1}^{-1}$ is the Kalman gain.

5) *Repeat steps 2)-4) until $t = T$.*

Note that in step 3) $\Omega_{t|t-1}x_1 = E(\alpha_t - \alpha_{t|t-1})(y_t - x_1'\alpha_{t|t-1})'$. Hence, updated estimates of α_t are computed using the least square projection of $\alpha_t - \alpha_{t|t-1}$ on $y_t - y_{t|t-1}$, multiplied by the prediction error. Similarly, $\Omega_{t|t-1} = E(\alpha_t - \alpha_{t|t-1})(\alpha_t - \alpha_{t|t-1})'$ is updated using a quadratic form involving the covariance between forecast errors in the two equations and the MSE of the forecasts. Note also that equations (6.7)-(6.8) provide the inputs for the next step of the recursion.

Example 6.4 *Consider extracting a signal α_t, for example, the long run trend of output, given that $\alpha_t = \alpha_{t-1}$ and that the trend is linked to output via $y_t = \alpha_t + v_{1t}$ where v_{1t} is a normal martingale difference process with variance $\sigma_{v_1}^2$. Using (6.6) we have that $\Omega_{t|t} = \Omega_{t|t-1} - \Omega_{t|t-1}(\Omega_{t|t-1} + \sigma_{v_1}^2)^{-1}\Omega_{t|t-1} = \frac{\Omega_{t|t-1}}{1+\frac{\Omega_{t|t-1}}{\sigma_{v_1}^2}} = \frac{\Omega_{t-1|t-1}}{1+\frac{\Omega_{t-1|t-1}}{\sigma_{v_1}^2}}$. Hence, starting from some $\Omega_0 = \bar{\Omega}_0$, we have $\Omega_{1|1} = \frac{\bar{\Omega}_0}{1+\frac{\bar{\Omega}_0}{\sigma_{v_1}^2}}$; $\Omega_{2|2} = \frac{\bar{\Omega}_0}{1+2\frac{\bar{\Omega}_0}{\sigma_{v_1}^2}}$; ...; $\Omega_{T|T} = \frac{\bar{\Omega}_0}{1+T\frac{\bar{\Omega}_0}{\sigma_{v_1}^2}}$. From (6.5) and (6.7), $\alpha_{T+1|T+1} = \alpha_{T|T} + \frac{\frac{\bar{\Omega}_0}{\sigma_{v_1}^2}}{1+T\frac{\bar{\Omega}_0}{\sigma_{v_1}^2}}(y_{T+1} - \alpha_{T|T})$. Hence, as $T \to \infty$, $\alpha_{T+1|T+1} = \alpha_{T|T}$ so that, asymptotically, the contribution of additional observations is negligible.*

Exercise 6.4 *Consider a vector MA process $y_t = e_t + e_{t-1}$ where $e_t \sim \mathbb{N}(0, I)$. Show that the optimal one-step ahead predictor for y_{t+1} is $y_{t+1|t} = \frac{t+1}{t+2}[y_t - y_{t|t-1}]$. Conclude that as $T \to \infty$, the optimal one-step ahead predictor is just last period's forecast error. (Hint: Cast the process into a state space format and apply the Kalman filter).*

Exercise 6.5 *Consider the process $y_t = A_1y_{t-1} + A_2y_{t-2} + e_t$. Here $\alpha_t = [y_t', y_{t-1}']'$, $v_{2t} = [e_t, 0]$, $\mathbb{D}_1 = \begin{bmatrix} \alpha_1 & \alpha_2 \\ 1 & 0 \end{bmatrix}$, $\Sigma_{v_2} = \begin{bmatrix} \sigma_e^2 & 0 \\ 0 & 0 \end{bmatrix}$, $\mathbb{D}_{0t} = v_{1t} = 0$, $x_1' = [1, 0]$. Show how to start the Kalman filter recursions; compute prediction and updated estimates of α_t for the first two observations.*

Exercise 6.6 *Suppose $y_{1t} = A_t y_{1t-1} + D_t y_{2t} + v_{1t}$ and $\alpha_t = (A_t, D_t) = \alpha_{t-1} + v_{2t}$, where y_{2t} are exogenous variables. Show the updating and prediction equations in this case. How would you handle the case of serially correlated v_{2t}?*

At times, it may be useful to construct estimates of the state vector which, at each t, contains information present in the entire sample. This is the case, in particular, in signal extraction problems; for example, when α_t is a common trend for a vector of y_t, we want estimates at each t to contain all the information available up to T. In this case the Kalman filter can be applied starting from the last observation, working backward through the sample, $t = T-1, \ldots, 1$, using $\alpha_{T|T}, \Omega_{T|T}$ and as initial conditions. That is:

$$\alpha_{t|T} = \alpha_{t|t} + (\Omega_{t|t}\mathbb{D}_1'\Omega_{t+1|t}^{-1})(\alpha_{t+1|T} - \mathbb{D}_1\alpha_{t|t}) \tag{6.9}$$

$$\Omega_{t|T} = \Omega_{t|t} - (\Omega_{t|t}\mathbb{D}_1'\Omega_{t+1|t}^{-1})(\Omega_{t+1|T} - \Omega_{t+1|t})(\Omega_{t|t}\mathbb{D}_1'\Omega_{t+1|t}^{-1})' \tag{6.10}$$

Equations (6.9)-(6.10) define the recursions of the so-called Kalman smoother.

Example 6.5 *Continuing with example 6.4, take $\alpha_{T|T}$ and $\Omega_{T|T}$ as initial conditions. Then $\Omega_{1|t} = \frac{\Omega_{T|T}}{1+T\frac{\Omega_{T|T}}{\sigma_{v_1}^2}}$ and $\alpha_{t|T} = \alpha_{t+1|T} + \frac{\frac{\Omega_{T|T}}{\sigma_{v_1}^2}}{1+T\frac{\Omega_{T|T}}{\sigma_{v_1}^2}}(y_{t|T} - \alpha_{t+1|T})$. Can you guess what $\alpha_{1|T}$ is?*

As a byproduct of the estimation, the Kalman filter allows us to transform (6.1)-(6.2) into a system driven by innovations in the measurement equation. In fact, using (6.5)-(6.7), it is immediate to see that (6.1) and (6.2) are equivalent to

$$y_t = x_{1t}'\alpha_{t|t-1} + \epsilon_t \tag{6.11}$$

$$\alpha_{t+1|t} = \mathbb{D}_1\alpha_{t|t} + \mathbb{D}_0 + \mathsf{K}_t\epsilon_t \tag{6.12}$$

where ϵ_t is the forecast error and $E_t(\epsilon_t\epsilon_t') \equiv \Sigma_{t|t-1}$. Hence, if the Kalman gain K_{t-1} is available and given $(\alpha_{1|0}, \Sigma_{1|0})$, $\alpha_{t|t-1}$ and ϵ_t can be computed recursively at any t. In turn, the Kalman gain is immediately obtained when $\Omega_{t-1|t-1}$ is available.

Exercise 6.7 *The reparametrization in (6.12)-(6.11) is trivial in the case of a constant coefficient VAR(q), since it is always possible to rewrite the measurement equation as $y_t = E[y_t|\mathcal{F}_{t-1}] + \epsilon_t$, where $\mathcal{F}_{t-1}$ is the information set at $t-1$. Show how to transform the ARMA(2,1) model of example 6.2 to fit such a representation.*

Hansen and Sargent (1998, pp.126-128) show that equation (6.6) can also be written as $\Omega_{t|t} = \mathbb{D}_1\Omega_{t-1|t-1}\mathbb{D}_1' + \mathbb{D}_2\Sigma_{v_2}\mathbb{D}_2' - \mathbb{D}_1\Omega_{t-1|t-1}x_1\Sigma_{t|t-1}^{-1}x_1'\Omega_{t-1|t-1}\mathbb{D}_1$. One can recognize in this expression a version of the matrix Riccati equation used in chapter 2 to solve linear regulator problems. Therefore, under regularity conditions, in state space models with constant coefficients, $\lim_{t\to\infty}\Omega_{t|t} = \Omega$. Consequently, $\lim_{t\to\infty}\mathsf{K}_t = \mathsf{K}$, and the stationary covariance matrix of the innovations is $\Sigma = \lim_{t\to\infty}\Sigma_{t|t} = x_1'\Omega x_1 + x_2'\Sigma_{v_1}x_2$. As we show next, the expressions for $\Omega, \mathsf{K}, \Sigma$ obtained in a constant coefficient model are the same as those asymptotically produced by a recursive least square estimator.

Example 6.6 *Consider estimating the constant (steady state) real interest rate α_t using T observations on the nominal interest rate y_t, demeaned by the average inflation rate, where $y_t = \alpha_t + v_{1t}$ and v_{1t} is a martingale difference process with variance $\sigma^2_{v_1}$. An unbiased minimum variance estimator is $\hat{\alpha}_T = \frac{1}{T}\sum_{t=1}^{T} y_t$. If y_{T+1} becomes available $\hat{\alpha}_{T+1} = \frac{1}{T+1}\sum_{t=1}^{T+1} y_t = \frac{T}{T+1}(\frac{1}{T}\sum_{t=1}^{T} y_t) + \frac{1}{T+1}y_{T+1} = \frac{T}{T+1}\hat{\alpha}_T + \frac{1}{T+1}y_{T+1}$ which is a recursive least square estimator. This estimator weights previous and current observations using the number of available observations and does not forget: each observation gets equal weight regardless of the time elapsed since it was observed. A more informative way to rewrite this expression is $\hat{\alpha}_{T+1} = \hat{\alpha}_T + \frac{1}{T+1}(y_{T+1} - \hat{\alpha}_T)$ and $\epsilon_t \equiv (y_{T+1} - \hat{\alpha}_T)$ is the innovation in forecasting y_{T+1}. Clearly, $\mathsf{K}_{T+1} = \frac{1}{T+1} \to 0$ as $T \to \infty$. Hence, as $T \to \infty$, $\hat{\alpha}_{T+1} \to \hat{\alpha}_T$.*

The recursions in (6.3)-(6.8) assume constant coefficients. The Kalman filter, however, can also be applied to models with time varying coefficients, as long as they are linear in parameters. For example, in the multivariate model

$$\begin{aligned} y_t &= \alpha_t y_{t-1} + v_{1t} \\ \alpha_t &= \alpha_{t-1} + v_{2t} \end{aligned} \tag{6.13}$$

recursive estimates of $\alpha_{t|t}$ and of the forecast error $\epsilon_t = y_t - \alpha_{t|t-1}y_{t-1}$ consistent with the information available at each t can be easily obtained. We extensively use models like (6.13) in chapter 10 when studying time varying Bayesian VAR models.

Exercise 6.8 *Consider the model $y_t = x_t'\alpha_t + v_{1t}$ where $\alpha_t = (I - \mathbb{D}_1)\alpha_0 + \mathbb{D}_1\alpha_{t-1} + v_{2t+1}$, α_0 is a constant; v_{1t} is a martingale difference with variance $\sigma^2_{v_1}$ and v_{2t} is a vector of martingale difference with variance Σ_{v_2}. Define $\alpha_t^\dagger = \alpha_t - \alpha_0$. Show the form of the updating equations for $\alpha_t^\dagger$ and Ω_t, assuming $\alpha_1^\dagger \sim \mathbb{N}(\alpha_{1|0}, \Omega_{1|0})$.*

A modified version of the Kalman filter can also be used in special nonlinear state space models; for example, those displaying structures like the one of exercise 6.3. To compute the Kalman gain in this case it is necessary to linearize the extended state space around the current estimate. For example, the updating equations are

$$\begin{aligned} \alpha_{t|t} &= \alpha_{t|t-1}\theta_{t|t-1} + \mathsf{K}_{1t}(y_t - \alpha_{t|t-1}) \\ \theta_{t|t} &= \theta_{t|t-1} + \mathsf{K}_{2t}(y_t - \alpha_{t|t-1}) \end{aligned} \tag{6.14}$$

where where $\mathsf{K}_{1t}, \mathsf{K}_{2t}$ are matrices involving linear and quadratic terms in the predictors $\theta_{t|t-1}$ and $\alpha_{t|t-1}$, linear terms in the variance $\sigma^2_{v_1}$ and in past Kalman gains (see Ljung and Soderstroem (1983), pp. 39-40 for details).

If initial conditions and innovations are normally distributed, the Kalman filter predictor is the best in both the class of linear and nonlinear predictors. Moreover, forecasts of y_t are normal with mean $x_1'\alpha_{t|t-1}$ and variance $\Sigma_{t|t-1}$. When the two above conditions are not satisfied, the Kalman filter only produces the best linear predictor for y_t, based on

information at time t. That is, there are nonlinear filters which produce more efficient estimators than those produced in (6.5)-(6.6). A nonlinear filter for a model with binomial innovations was described in chapter 3 (see also Hamilton (1994), ch.22).

Example 6.7 *As we have seen, a two-state Markov switching model for y_t can be written as $y_t = a_0 + a_1\varkappa_t + y_{t-1}$ where $\varkappa_t$ has an AR(1) representation of the form*

$$\varkappa_t = (1 - p_2) + (p_1 + p_2 - 1)\varkappa_{t-1} + v_{1t} \tag{6.15}$$

and where v_{1t} can take four possible values $[1 - p_1, -p_1, -(1 - p_2), p_2]$ with probabilities $[p_1, 1-p_1, p_2, 1-p_2]$ and therefore is non-normal. It is immediate to verify that this process has a state space representation and that the orthogonality assumptions needed for identification are satisfied. However, while $corr(v_{1t}, \varkappa_{t-\tau}) = 0\ \forall\tau > 0$, the two processes are not independent. Equation (6.15) can be rewritten as

$$(1-(p_1+p_2-1)\ell)\Delta y_t = a_1(1-(p_1+p_2-1)\ell)\varkappa_t = a_1(1-p_2)+a_0(2-p_1-p_2)+v_{1t} \tag{6.16}$$

Hence, although y_t has a linear ARIMA(1,1,0) structure, Kalman filter estimates of $y_{t+1|t}$ based on such a model are suboptimal since the non-linear structure present in v_{1t} is ignored. In fact, optimal forecasts are obtained using

$$E_t\Delta y_{t+1} = a_0 + a_1 E_t\varkappa_{t+1} = a_0 + a_1[\frac{1-p_2}{2-p_1-p_2} + (p_1+p_2-1)(P[\varkappa_t = 1|\mathcal{F}_t] - \frac{1-p_2}{2-p_1-p_2})] \tag{6.17}$$

where $\mathcal{F}_t$ represents the information set at t. The nonlinear filtering algorithm described in chapter 3 uses (6.17) to obtain estimates of $\varkappa_t$.

While we have assumed that the measurement error and the error in the state equation are uncorrelated, in some situations this assumption may be unpalatable. For example, in the context of a model like (6.13), one may want to have the innovations in y_t and in α_t to be correlated. Relaxing this assumption requires some ingenuity. The next exercise shows that a system with a serially correlated measurement error is equivalent to a system with correlation between innovations in the transition and the measurement equations.

Exercise 6.9 *Suppose that all coefficients are constant, that $\mathbb{D}_0 = 0$ and that v_t in equation (6.1) satisfies $v_{1t} = \rho_v v_{1t-1} + \upsilon_t$ where ρ_v has all the eigenvalues less than one in absolute value and υ_t is a martingale difference with covariance matrix Σ_υ. Assuming that $E(v_{2t}\upsilon'_\tau) = 0\ \forall t$, and $\tau \neq t$, show that an equivalent state space representation is given by (6.2) and by $y_t^\dagger = x_{1t}^\dagger\alpha_t + v_{1t+1}^\dagger$ where $y_t^\dagger = y_{t+1} - \rho_v y_t$, $x_{1t}^\dagger = x_{1t}\mathbb{D}_1 - \rho_v x_{1t}$ and $v_{1t+1}^\dagger = x_{1t}\mathbb{D}_2 v_{2t+1} + v_{1t+1}$.*

Exercise 6.10 *Suppose α_t is normally distributed with mean $\bar{\alpha}$ and variance $\bar{\Sigma}_\alpha$, that $y_t = x_1'\alpha_t + v_{1t}$, where v_{1t} is orthogonal to α_t, and $v_{1t} \sim iid\ \mathbb{N}(0, \sigma_{v_1})$.*
(i) Show that $y_t \sim \mathbb{N}(x_1'\bar{\alpha}, x_1'\bar{\Sigma}_\alpha x_1 + \sigma_{v_1}^2)$.

(ii) Using the fact that the posterior density of α_t *is* $g(\alpha_t|y_t) = \frac{g(\alpha_t)f(y_t|\alpha_t)}{g(y_t)}$*, show that* $g(\alpha_t|y_t) \propto exp\{-0.5((\alpha_t - \bar{\alpha})'\bar{\Sigma}_\alpha^{-1}(\alpha_t - \bar{\alpha}) + (y_t - x_1'\alpha_t)'\sigma_{v_1}^{-2}(y_t - x_1'\alpha_t)\} \equiv exp\{-0.5((\alpha_t - \tilde{\alpha})'\tilde{\Sigma}_\alpha^{-1}(\alpha_t - \tilde{\alpha})\}$ *where* $\tilde{\alpha} = \bar{\alpha} + \bar{\Sigma}_\alpha x_1\sigma_{v_1}^{-2}(y_t - x_1'\bar{\alpha})$ *and* $\tilde{\Sigma}_\alpha = \bar{\Sigma}_\alpha + \bar{\Sigma}_\alpha x_1\sigma_{v_1}^{-2}x_1'\bar{\Sigma}_\alpha$*.*

Exercise 6.11 *A generalized version of a log-linearized RBC model can be written as* $\alpha_t = \mathbb{D}_{1t-1}\alpha_{t-1} + v_{2t},\ v_{2t} \sim (0, \Sigma_t)$*, and* $y_t = x_{1t}'\alpha_t$ *where* α_t *represents a vector of states and shocks and* y_t *are the controls. Assume that* $\Sigma_t, x_{1t}, \mathbb{D}_{1t-1}$ *are known.*
(i) Find the updating equation for the forecast error variance and show that $x_{1t}'\Omega_{t|t}x_{1t} = 0$*.*
(ii) Show that $\Omega_{t+1|t} = \mathbb{D}_{1t}\Omega_{t|t}\mathbb{D}_{1t}' + \Sigma_t$*.*

Given the recursive nature of Kalman filter estimates, it is easy to compute multistep forecasts of y_t. We leave the derivation of these forecasts as an exercise for the reader.

Exercise 6.12 *Consider the model (6.1)-(6.2) and the prediction of* $y_{t+\tau}$*. Show that the* τ*-steps ahead forecast error is* $x_{1t+\tau}'(\alpha_{t+\tau} - \alpha_{t+\tau,t}) + x_{2t+\tau}'v_{1t+\tau}$ *and that the MSE of the forecast is* $x_{1t+\tau}'\Omega_{t+\tau|t}x_{1t+\tau} + x_{2t+\tau}'\Sigma_{v_1}x_{2t+\tau}$*. Show the form of* $\alpha_{t+\tau|t}$ *and* $\Omega_{t+\tau|t}$*.*

Example 6.8 *Consider an* $m \times 1$ *VAR(q) model,* $\mathbb{Y}_t = \mathbb{A}\mathbb{Y}_{t-1} + E_t$*. As we have seen in example 6.1, this is a state space model for* $x_{1t}' = I,\ \alpha_t = y_t,\ \mathbb{D}_{1t} = \mathbb{A},\ \Sigma_{v_1} = 0,\ v_{2t} = E_t,\ \mathbb{D}_{2t} = I, \mathbb{D}_{0t} = 0$*. The* τ*-steps ahead forecast of* y_t *is* $E_t[y_{t+\tau}] = \mathbb{S}\mathbb{A}^\tau\mathbb{Y}_t$*, where* $\mathbb{S}$ *is a selection matrix. The forecast error variance is* $(\mathbb{S}(\mathbb{Y}_{t+\tau} - \mathbb{A}^\tau\mathbb{Y}_t))(\mathbb{S}(\mathbb{Y}_{t+\tau} - \mathbb{A}^\tau\mathbb{Y}_t)'))$*.*

6.2 The Prediction error decomposition of likelihood

Maximum likelihood estimation of nonlinear models is complicated. However, even in models like (6.1)-(6.2), which are conditionally linear in the parameters, maximization of the likelihood function is problematic when observations are not independent. This section is concerned with the practical question of constructing the likelihood function for models which have a format like (6.1)-(6.2), when y_t is serially correlated over time. It turns out that there is a convenient format, called prediction error decomposition, which can be used to estimate ARMA, structural VARs and, as we will see, DSGE models.

To understand what this decomposition entitles let $f(y_1, \dots y_T)$ be the joint density of $\{y_t\}_{t=1}^T$. Given the properties of joint densities, it is possible to decompose $f(y_1, \dots y_T)$ into the product of a conditional and a marginal, and repeatedly substituting we have:

$$\begin{aligned} f(y_1, \dots y_T) &= f(y_T|y_{T-1} \dots y_1)f(y_{T-1}, \dots y_1) \\ &= f(y_T|y_{T-1} \dots y_t)f(y_{T-1}|y_{T-2}, \dots y_1)f(y_{T-2}, \dots y_1) \\ &\dots \\ &= \prod_{j=0}^{J} f(y_{T-j}|y_{T-j-1} \dots y_1)f(y_1) \end{aligned} \tag{6.18}$$

and $\log f(y_1, \dots y_T) = \sum_j \log f(y_{T-j}, |y_{T-j-1} \dots y_1) + \log f(y_1)$. If $y = [y_1, \dots y_T] \sim \mathbb{N}(\bar{y}, \Sigma_y)$

$$\mathcal{L}(y|\phi) = \log f(y_1, \dots, y_T|\phi) = -\frac{T}{2}(\log 2\pi + \log|\Sigma_y|) - \frac{1}{2}(y-\bar{y})\Sigma_y^{-1}(y-\bar{y}) \tag{6.19}$$

where $\phi = (\bar{y}, \Sigma_y)$. Calculation of (6.19) requires the inversion of Σ_y, which is a $T \times T$ matrix, and this may be complicated when T is large. Using decomposition (6.18), we can partition $\mathcal{L}(y_1, \dots, y_t|\phi) = \mathcal{L}(y_1, \dots, y_{T-1}|\phi)\mathcal{L}(y_t|y_{T-1}, \dots, y_1, \phi)$. When$\{y_t\}_{t=1}^T$ is normal, both the conditional and the marginal blocks are normal.

Let $y_{t|t-1}$ be a predictor of y_t using information up to $t-1$. The prediction error is $\epsilon_t = y_t - y_{t|t-1} = y_t - E(y_t|y_{t-1}, \dots, y_1) + E(y_t|y_{t-1}, \dots, y_1) - y_{t|t-1}$ and its Mean Square Error (MSE) is $E(\epsilon_t - E(\epsilon_t))^2 = E(y_t - E(y_t|y_{t-1}, \dots y_1))^2 + E(E(y_t|y_{t-1}, \dots y_1) - y_{t|t-1})^2$. The best predictor of y_t, i.e. the one that makes the MSE of the prediction error as small as possible, is obtained when $E(y_t|y_{t-1}, \dots, y_1) = y_{t|t-1}$. Given this choice, the MSE of ϵ_t, denoted by $\sigma^2_{\epsilon_t}$, equals the variance of $(y_t|y_{t-1}, \dots, y_1)$.

The conditional density of y_t given information at time $t-1$ can then be written as:

$$\mathcal{L}(y_t|y_{t-1}, \dots, y_1, \sigma^2_{\epsilon_t}) = -\frac{1}{2}\log(2\pi) - \log(\sigma_{\epsilon_t}) - \frac{1}{2}\frac{(y_t - y_{t|t-1})^2}{\sigma^2_{\epsilon_t}} \tag{6.20}$$

Since (6.20) is valid for any $t > 1$ using (6.18) we have that

$$\begin{aligned}
\mathcal{L}(y_1, \dots y_T|\sigma^2_{\epsilon_1}, \dots, \sigma^2_{\epsilon_T}) &= \sum_{t=2}^{T} \mathcal{L}(y_t|y_{t-1}, \dots, y_1, \sigma^2_{\epsilon_2}, \dots, \sigma^2_{epsilon_T}) + \mathcal{L}(y_1|\sigma^2_{\epsilon_1}) \\
&= -\frac{T-1}{2}\log(2\pi) - \sum_{t=2}^{T}\log\sigma_{\epsilon_t} - \frac{1}{2}\sum_{t=2}^{T}\frac{(y_t - y_{t|t-1})^2}{\sigma^2_{\epsilon_t}} \\
&- \frac{1}{2}\log(2\pi) - \log\sigma_{\epsilon_1} - \frac{1}{2}\frac{(y_1 - \bar{y}_1)^2}{\sigma^2_{\epsilon_1}}
\end{aligned} \tag{6.21}$$

where $\bar{y}_1$ is the unconditional predictor of y_1. (6.21) is the decomposition we were looking for. Three important aspects need to be emphasized. First, (6.21) can be computed recursively, since it only involves one step ahead prediction errors and their optimal MSE. This should be contrasted with (6.19) where the entire vector of y_t's is used. Second, both the best predictor $y_{t|t-1}$ and the MSE of the forecast $\sigma^2_{\epsilon_t}$ vary with time. Therefore, we have transformed a time invariant problem into a problem involving quantities that vary over time. Third, if y_1 is a constant, prediction errors are constant and exactly equal to the innovations in y_t.

Example 6.9 *Consider a univariate AR(1) process $y_t = Ay_{t-1} + e_t$, $|A| < 1$, where e_t is normal martingale difference process with variance σ^2_e. Let $\phi = (A, \sigma^2_e)$. Assume that the process has started far in the past but it has been observed only from $t = 1$ on. For any t, $y_{t|t-1} \sim \mathbb{N}(Ay_{t-1}, \sigma^2_e)$. Hence, the prediction error $\epsilon_t = y_t - y_{t|t-1} = y_t - Ay_{t-1} = e_t$.*

Moreover, since the variance of e_t is constant, also the variance of the prediction error is constant (from time $t = 2$ on). Setting $\bar{y} = 0$, $y_1 \sim \mathbb{N}(0, \frac{\sigma_e^2}{1-A^2})$ and

$$\begin{aligned}\mathcal{L}(\phi) &= \sum_{t=2}^{T} \mathcal{L}(y_t|y_{t-1}, \ldots, y_1, \phi) + \mathcal{L}(y_1|\phi) \\ &= -\frac{T}{2}\log(2\pi) - T\log(\sigma_e) - \frac{1}{2}\sum_{t=2}^{T}\frac{(y_t - Ay_{t-1})^2}{\sigma_e^2} + \frac{1}{2}(\log(1-A^2) - \frac{(1-A^2)y_1^2}{\sigma_e^2})\end{aligned}$$

Hence $\sigma_{\epsilon_t}^2 = \sigma_e^2$ for all $t \geq 2$, while $\sigma_{\epsilon_1}^2 = \frac{\sigma_e^2}{1-A^2}$.

Exercise 6.13 *Consider the univariate model $y_{1t} = A_1(\ell)y_{1t-1} + D(\ell)e_t + A_2 y_{2t}$, where y_{2t} are exogenous variables, $A_1(\ell)$ is a polynomial of order q_1, $D(\ell)$ is a polynomial of order q_2. Find $y_{1t|t-1}$ and $\sigma_{\epsilon_t}^2$ in this case. Show the form of the log likelihood function assuming that the first $q = max[q_1, q_2 + 1]$ values of $y_t = [y_{1t}, y_{2t}]$ are constants.*

Taking the initial observations as given is convenient since it eliminates a source of nonlinearities. In general, nonlinearities do not allow to compute an analytical solution to the first order conditions of the maximization problem and the maximum of the likelihood must be located using numerical techniques. Conditioning on the initial observations makes the maximization problem trivial in many cases. Note also that, as $T \to \infty$, the contribution of the first observation to the likelihood becomes negligible. Therefore, exact and conditional maximum likelihood coincide if the sample is large. Furthermore, when the model has constant coefficients, the errors are normally distributed and the initial observations fixed, maximum likelihood and OLS estimators are identical (see chapter 4 in the case of a VAR). This would not be the case when a model features moving average terms (see example 6.11), since nonlinearities do not wash out, even conditioning on the initial observations.

Example 6.10 *Consider finding the ML estimator of the AR process described in example 6.9. Conditioning on y_1 the log likelihood of $(y_2, \ldots, y_T)$ is proportional to $\sum_{t=2}^{T}\{-\log(\sigma_e) - \frac{1}{2\sigma_e^2}(y_t - Ay_{t-1})^2\}$. Maximizing this quantity with respect to A (conditional on σ_e^2), is equivalent to minimizing $(y_t - Ay_{t-1})^2$, which produces $A_{ML} = A_{ols}$. Using A_{ML}, the likelihood can be concentrated to obtain $-\frac{T-1}{2}\log(\sigma_e^2) - \frac{\sum_t \epsilon_t'\epsilon_t}{2\sigma_e^2}$. Maximizing it with respect to σ_e^2 leads to $\sigma_{ML}^2 = \frac{\sum_t \epsilon_t'\epsilon_t}{T-1}$. Suppose now that we do not wish to condition on y_1. Then the likelihood function is proportional to $\sum_{t=2}^{T}\{-\log(\sigma_e) - \frac{1}{2\sigma_e^2}(y_t - Ay_{t-1})^2\} + \{-0.5\log(\frac{\sigma_e^2}{1-A^2}) - \frac{y_1^2(1-A^2)}{2\sigma_e^2}\}$. If $T \to \infty$, the first observation makes a negligible contribution to the likelihood of the sample. Therefore, conditional ML estimates of A asymptotically coincide with full ML estimates, provided $|A| < 1$.*

Consider, finally, the case where A is time varying, e.g. $A_t = \mathbb{D}_1 A_{t-1} + v_{2t}$. Conditional on some A_0, the recursive conditional maximum likelihood estimator of $A_{t|t}$ and the smoothed maximum likelihood estimator $A_{t|T}$ can be obtained with the Kalman filter and the Kalman smoother. As $T \to \infty$, the importance of the initial observation will be discounted as long as the roots of $\mathbb{D}_1$ are all less than one in absolute value.

Exercise 6.14 *(i) Suppose that $y_t = x_t'\alpha + e_t$ where e_t is normal martingale difference with variance $\sigma^2_{e_t}$ and let x_t be fixed regressors. Show how to derive the prediction error decomposition of the likelihood for this model.*
(ii) Let x_t be a random variable, normally distributed with mean $\bar{x}$ and variance Σ_x. Show how to compute the prediction error decomposition of the likelihood in this case.

Multivariate prediction error decompositions present no difficulties. If y_t is $m \times 1$ vector

$$\mathcal{L}(y|\phi) = -\frac{Tm}{2}\log(2\pi) - \frac{1}{2}\sum_{t=1}^{T}\log|\Sigma_{t|t-1}| - \frac{1}{2}\sum_{t=1}^{T}(y_t - y_{t|t-1})\Sigma^{-1}_{t|t-1}(y_t - y_{t|t-1}) \quad (6.22)$$

where $\epsilon_t = y_t - y_{t|t-1} \sim \mathbb{N}(0, \Sigma_{t|t-1})$ and where we assume $y_1 \sim \mathbb{N}(\bar{y}_1, \Sigma_{1|0})$ and $\epsilon_1 = y_1 - \bar{y}_1$.

Exercise 6.15 *Consider the setup of exercise 6.11. Show the form of $y_{t|t-1}$ and $\Sigma_{t|t-1}$ and the prediction error decomposition of the likelihood in this case.*

The prediction error decomposition is convenient in two respects. First, the building blocks of the decomposition are the forecast errors ϵ_t and their MSE $\Sigma_{t|t-1}$. Since the Kalman filter produces these quantities recursively, it can be used to build the prediction error decomposition of the likelihood of any model which has a state space format. Second, since any ARMA process has a state space format, the prediction error decomposition of the likelihood can be easily obtained for a variety of statistical an economic models.

Maximization of the likelihood conditional on the initial observations, can be obtained by extending algorithm 6.1. Let $\phi = [vec(x_1'), vec(x_2'), vec(\mathbb{D}_1), vec(\mathbb{D}_0), vec(\mathbb{D}_2), \Sigma_{v_1}, \Sigma_{v_2}]$. Then

Algorithm 6.2

1) *Choose some initial $\phi = \phi_0$.*

2) *Do steps 1)-4) of algorithm 6.1.*

3) *At each step save $\epsilon_t = y_t - y_{t|t-1}$ and $\Sigma_{t|t-1}$. Construct the log likelihood (6.22).*

4) *Update initial estimates of ϕ using any of the methods described in section 6.3.*

5) *Repeat steps 2)-4) until $|\phi^l - \phi^{l-1}| \leq \iota$; or $\frac{\partial \mathcal{L}(\phi)}{\partial \phi}|_{\phi=\phi^l} < \iota$, or both, for ι small.*

Two comments on algorithm 6.2 are in order. First, the initial values of iterations can be typically obtained by running an OLS regression on the constant coefficient version of the model. If the assumptions underlying the state space specification are correct this will consistently estimate the average value of the parameters. Second, for large dimensional problems, maximization routines typically work better if Choleski factor of $\Sigma_{t|t-1}$, is used in the computations of the likelihood.

The conditional prediction error decomposition is particularly useful to estimate models with MA terms. Such models are difficult to deal with in standard setups but fairly easy to estimate within a state space framework.

Example 6.11 *In testing the efficiency of foreign exchange markets one runs a monthly regression of the realized three month change in spot exchange rate at $t+3$ on the forward premium quoted at t for $t+3$. As seen in chapter 5, such a regression has moving average errors of order up to 2 because of overlapping time intervals. Therefore, a model for testing efficiency could be $y_{t+3} = b_0 x_t + \epsilon_{t+3}$ with $\epsilon_{t+3} = e_{t+3} + b_1 e_{t+2} + b_2 e_{t+1}$ where, under the null hypothesis, $b_0 = 1$ and e_t is a normal martingale difference with variance σ_e^2. This model can be cast into a state space framework by defining $\mathbb{D}_0 = 0, \mathbb{D}_2 = I, x'_{2t} = I, v_{1t} = 0$,* $\alpha_t = \begin{bmatrix} x_t \\ e_{t+3} \\ e_{t+2} \\ e_{t+1} \end{bmatrix}$, $\mathbb{D}_1 = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}$, $x_{1t} = \begin{bmatrix} b_0 \\ 1 \\ b_1 \\ b_2 \end{bmatrix}$ $v_{2t} = \begin{bmatrix} x_t \\ e_{t+3} \\ 0 \\ 0 \end{bmatrix}$. *Suppose we are interested in estimating $[b_0, b_1, b_2]$ and in testing $b_0 = 1$. Then ML estimates can be obtained starting the Kalman filter at $\alpha_{1|0} = [x_1, 0, 0, 0]$ and $\Omega_{1|0} = diag\{\sigma_x^2, \sigma_e^2, \sigma_e^2, \sigma_e^2\}$ where σ_x^2 is the unconditional variance of the forward premium and σ_e^2 could be either the variance of the error $\hat{e}_t = y_t - \hat{b}_0 x_{t-3}$ in a training sample (say, from $-\tau$ to 0) or set to an arbitrarily large number. To start the iterations we need x_{10}, that is, we need some initial estimates of (b_0, b_1, b_2). An estimate of b_0 could be obtained in a training sample or, if no such a sample exists, using available data but disregarding serial correlation in the error term. Initial estimates of b_1 and b_2 could then be $b_1 = b_2 = 0$. Then the sequence of iterations producing $\alpha_{t|t-1}$ and $\Omega_{t|t-1}$ can be used to compute the likelihood function. Note that for this simple problem one could evaluate the likelihood numerically at successive grids of, say, 20 points in each dimension and locate the maximum numerically.*

Exercise 6.16 *Consider an AR(2) process $y_t = A_0 + A_1 y_{t-1} + A_2 y_{t-2} + e_t$ where $e_t \sim iid\ \mathbb{N}(0, \sigma_e^2)$. Show that the exact log likelihood function is $\mathcal{L}(\phi) \propto -T\log(\sigma_e) + 0.5log((1 + A_2)^2[(1 - A_2)^2 - A_1^2]) - \frac{1+A_2}{2\sigma^2}[(1 - A_2)(y_1 - \bar{y})^2 - 2A_1(y_1 - \bar{y})(y_2 - \bar{y}) + (1 - A_2)(y_2 - \bar{y})^2] - \sum_{t=3}^{T} \frac{(y_t - A_0 - A_1 y_{t-1} + A_2 y_{t-2})^2}{2\sigma^2}$ where $\bar{y} = \frac{A_0}{1-A_1-A_2}$. Which terms disappear if a conditional likelihood approach is used? Show that $\sigma_{ML}^2 = \frac{1}{T-2}\sum_{t=3}^{T}(y_t - A_{0,ML} - A_{1,ML} y_{t-1} - A_{2,ML} y_{t-2})^2$.*

6.2.1 Some Asymptotics of ML estimators

It is fairly standard to show that, under regularity conditions, ML estimates of the parameters of a state space model are consistent and asymptotically normal (see e.g. Harvey (1991)). The conditions needed are generally of three types. First, we need the state equation to define a covariance stationary process. One simple sufficient condition for this is that the eigenvalues of $\mathbb{D}_{1t}$ are all less than one in absolute value for all t. Second, if the model includes exogenous variables we also need them to be covariance stationary, linearly regular processes. Third, we need the true parameters not to lie on the boundary of the parameter space. Then, under the above conditions, $\sqrt{T}(\phi_{ML} - \phi_0) \xrightarrow{D} \mathbb{N}(0, \Sigma_\phi)$ where $\Sigma_\phi = -T^{-1}E(\sum_t \frac{\partial^2 \mathcal{L}}{\partial\phi\partial\phi'}|_{\phi=\phi_0})^{-1}$.

For the case in which the innovations are the errors in the measurement equation, the asymptotic covariance matrix is block diagonal, as it is shown next.

Example 6.12 *For an AR(1) model it is quite easy to derive Σ_ϕ. In fact, conditional on the initial observations, the log likelihood is $\mathcal{L}(\phi) \propto -\frac{T-1}{2}\log\sigma_\epsilon^2 - \frac{\sum_{t=2}^{T}\epsilon_t^2}{2\sigma_\epsilon^2}$ where $\epsilon_t = y_t - Ay_{t-1}$ and the matrix of second derivatives is $\begin{bmatrix} -\sigma_\epsilon^{-2}\sum_t y_{t-1}^2 & -\sigma_\epsilon^{-4}\sum_t \epsilon_t y_{t-1} \\ -\sigma_\epsilon^{-4}\sum_t \epsilon_t y_{t-1} & (2\sigma_\epsilon^4)^{-1}(T-1) - \sigma_\epsilon^{-6}\sum_t \epsilon_t^2 \end{bmatrix}$. Since the expectation of the off-diagonal elements is zero, the asymptotic covariance matrix is diagonal with $var(A) = \frac{\sigma_\epsilon^2}{(T-1)\sum_t y_{t-1}^2}$ and $var(\sigma_e^2) = \frac{2\sigma_\epsilon^4}{T-1}$.*

The derivation of the Kalman filter assumes that the innovations in the measurement and in the observation equations are normally distributed. Since the likelihood function is calculated with the Kalman filter estimates, one may wonder what are the properties of maximum likelihood estimates when the distribution of the driving forces is misspecified.

As mentioned, misspecification of the distribution of the errors does not create consistency problems for Kalman filter estimates. It turns out that this property carries over to maximum likelihood estimates. In fact, maximum likelihood estimates obtained incorrectly assuming a normal distribution (typically called quasi-ML) have nice properties under a set of regularity conditions. We ask the reader to verify that this is the case for a simple problem in the next exercise.

Exercise 6.17 *Suppose observations on y_t are drawn from a t-distribution with a small number of degrees of freedom (say, less than 5) but that an econometrician estimates the constant coefficient state space model $y_t = \alpha_t + v_{1t}$, $\alpha_t = \alpha_{t-1}$ where v_{1t} is a normal martingale difference with variance $\sigma_{v_1}^2$. Show that the ML estimator for α_t based on the wrong (normal) distribution will be consistent and asymptotically normal. Show the form of the asymptotic covariance matrix.*

Intuitively, if the sample size is large and homogeneous, a normal approximation is appropriate. In the context of a constant coefficient state space model, we could have achieved the same conclusion by noting that recursive OLS is consistent and asymptotically normal if the regressors are stationary, ergodic and uncorrelated with the errors and that recursive OLS and Kalman filter-ML estimates coincide if a conditional likelihood is used.

When the coefficients of the state space model are time varying, ML estimates obtained with misspecified errors are no longer asymptotically equivalent to those of the correct model and Kalman filter estimates are not best linear MSE estimates of α_t.

We have seen that maximum likelihood estimates have an asymptotic covariance matrix equal to which is $-\frac{1}{T}E(\frac{\partial^2\mathcal{L}(\phi)}{\partial\phi\partial\phi'}|_{\phi=\phi_0})^{-1}$. There are many ways to estimate this matrix. One is to evaluate the quantity at the ML estimator, substituting averages for expectations, that is, $\text{var}_1(\phi) = (-\sum_t \frac{\partial^2\mathcal{L}_t(\phi)}{\partial\phi\partial\phi'}|_{\phi=\phi_{ML}})^{-1}$. An alternative is obtained noting that an approximation to the second derivatives of the likelihood function can be calculated taking the derivatives of the scores, i.e. $\text{var}_2(\phi) = (\sum_t(\frac{\partial\mathcal{L}_t(\phi)}{\partial\phi}|_{\phi=\phi_{ML}})(\frac{\partial\mathcal{L}_t(\phi)}{\partial\phi}|_{\phi=\phi_{ML}})')^{-1}$.

Finally, a quasi ML estimator can be obtained combining the two above estimators. That is, $\text{var}_3(\phi) = -((var_1(\phi))(var_2(\phi))^{-1}(var_1(\phi))$.

Exercise 6.18 *For the AR(1) model considered in example 6.12, show the form of the three estimates of the asymptotic covariance matrix.*

Hypothesis testing on the parameters is fairly standard. Given the asymptotic normality of ML estimates, one could use t-tests to verify simple restrictions on the parameters or likelihood ratio tests when more general hypotheses are involved.

Example 6.13 *Continuing with example 6.11, to test $b_0 = 1$ use $\frac{b_{0,ML}-1}{\sigma_{b_{0,ML}}}$ and compare it to a t distribution with $T-1$ degrees of freedoms (or to a normal (0,1), if T is large). Alternatively, one could estimate the model under the restriction $b_0 = 1$, construct the likelihood function, calculate $-2[\mathcal{L}(b_{0,ML}) - \mathcal{L}(b_0 = 1)]$ and compare it with a $\chi^2(1)$.*

As we have seen with GMM, it may be more convenient at times to use estimates of a restricted model. This would be the case, for example, if the model is non-linear, but it becomes linear under some restrictions, or if it contains MA terms. In this case one can use the Lagrangian Multiplier (LM) statistic $\frac{1}{T}[\sum_t(\frac{\partial\mathcal{L}(\phi)}{\partial\phi})|_{\phi^{re}}]'\Sigma_\phi^{-1}[(\sum_t \frac{\partial\mathcal{L}(\phi)}{\partial\phi})_{\phi^{re}})] \sim \chi^2(\nu)$, where ν is the number of restrictions.

Example 6.14 *For the model of example 6.2, if $D_1 = 0$, conditional ML estimates of $A = [A_1, A_2]'$ solve the normal equations $Ax'x = x'y$ where $x_t = [y_{t-1}, y_{t-2}]$, $x = [x_1, \dots x_t]'$. However, if $D_1 \neq 0$ the normal equations are nonlinear and no analytical solution exists. Therefore, one may impose $D_1 = 0$ for estimation and test if the restriction holds.*

Two non-nested hypotheses can be evaluated using, for example, forecasting accuracy tests like the one of Diebold and Mariano (1995). Let ϵ_t^i be the prediction errors produced by specification $i = 1, 2$ and let $h_t = (\epsilon_t^1)^2 - (\epsilon_t^2)^2$. Then, under the hypothesis of similar predictive accuracy, the statistic $\mathsf{S} = \frac{\bar{h}}{se(h)}$, where $\bar{h} = \frac{1}{T}\sum_t h_t$, $se(h) = \sqrt{\frac{1}{T}\sum_t(h_t - \bar{h})^2}$ is asymptotically normally distributed with mean zero and variance one. We will use this statistic in section 6.5 when comparing the forecasting accuracy of a DSGE model relative to an unrestricted VAR.

6.3 Numerical tips

There are many ways to update initial estimates in step 4) of algorithm 6.2. Here we only briefly list some of them and highlighting advantages and disadvantages of each.

- Grid search.
  This maximization method is feasible when the dimension of ϕ is small. It involves discretizing the problem and selecting the value of ϕ which achieves the maximum on the grid. One advantage of the approach is that no derivatives of the likelihood are

needed - which can be useful if the problem is complicated. When the likelihood is globally concave, the approach will find an approximation to the maximum. However, if multiple peaks are present, it may select local maxima. For this reason, the grid should be fine enough to avoid pathologies. While care should be exercised in taking them as final estimates, grid estimates are useful as initial conditions for other algorithms.

- Simplex method
  A k-dimensional simplex is spanned by $k+1$ vectors which are the vertices of the simplex (e.g. if $k = 2$, two dimensional simplexes are triangles). This method is typically fast and works as follows. If a maximum is found at some iteration, the method substitutes it with a point on the ray from the maximum through the centroid of the remaining points. For example, if $\mathcal{L}(\phi_m) = \max_{j=1,k+1} \mathcal{L}(\phi_j)$, we replace ϕ_m by $\varrho\phi_m + (1-\varrho)\bar{\phi}$, where $\bar{\phi}$ is the centroid, $0 < \varrho < 1$ and repeat the maximization. This approach does not require the calculation of gradients or second derivatives of the likelihood and can be used when other routines fail. The major disadvantage is that no standard errors for the estimates are available.

- Gradient methods
  All algorithms in this class update initial estimates by taking a step based on the gradient of the likelihood at the initial estimate. They differ in the size and the direction in which the step is taken.

  a) Method of Steepest ascent.
  At each iteration l, parameters are updated using: $\phi^l = \phi^{l-1} + \frac{1}{2\lambda} gr(\phi^l)$ where $gr(\phi^l) = \frac{\partial \mathcal{L}(\phi)}{\partial \phi}|_{\phi=\phi^l}$ and λ is the Lagrangian multiplier of the problem $\max_{\phi^l} \mathcal{L}(\phi^l)$ subject to $(\phi^l - \phi^{l-1})'(\phi^l - \phi^{l-1}) = \kappa$. In words, the method updates current estimates using the scaled gradient of the likelihood. λ is a smoothness parameter which prevents large jumps in ϕ between iterations (it plays the same role as λ in the Hodrick Prescott or exponential smoothing filters). Note that if $\phi^l \approx \phi^{l-1}$, $gr(\phi^l) \approx gr(\phi^{l-1})$ and one can use $\phi^l = \phi^{l-1} + \varrho gr(\phi^{l-1})$ where ϱ is small positive scalar (e.g. 10^{-5}). This choice is very conservative and avoids jumps in the estimates. However, a lot of iterations are typically needed before convergence is achieved and convergence could only be to local maximum. It is therefore a good idea to start the algorithm from several initial conditions and check whether the same maximum is obtained.

  b) Newton-Raphson Method
  The method is applicable if $\frac{\partial^2 \mathcal{L}(\phi)}{\partial \phi \partial \phi'}$ exists and if $\mathcal{L}(\phi)$ is concave (i.e. the matrix of second derivatives is positive definite). In this case, taking a second order expansion of $\mathcal{L}(\phi)$ around ϕ_0, we have:

$$\mathcal{L}(\phi) = \mathcal{L}(\phi_0) + gr(\phi_0)(\phi - \phi_0) - 0.5(\phi - \phi_0)' \frac{\partial^2 \mathcal{L}(\phi)}{\partial \phi \partial \phi'} (\phi - \phi_0) \tag{6.23}$$

Maximizing (6.23) with respect to ϕ and using ϕ^{l-1} as an estimate of ϕ_0 we have

$$\phi^l = \phi^{l-1} + (\frac{\partial^2 \mathcal{L}(\phi)}{\partial\phi\partial\phi'}|_{\phi=\phi^{l-1}})^{-1} gr(\phi^{l-1}) \tag{6.24}$$

If likelihood is quadratic, (6.24) generates convergence in one step. If it is close to quadratic, iterations on (6.24) will converge quickly and the global maximum will be achieved. However, if the likelihood is far from quadratic, not globally concave or if ϕ_0 is far away from the maximum, the method may have worse properties than the method of steepest ascent. Note that $(\frac{\partial^2 \mathcal{L}(\phi)}{\partial\phi\partial\phi'})^{-1}$ can be used to provide an estimate of the variance covariance matrix of ϕ at each iteration. Once could combine steepest-ascent and Newton-Raphson methods into an hybrid one which shares the good properties of both, may speed up calculation without producing large jumps in the parameters estimates. This is done, e.g., by choosing $\phi^l = \phi^{l-1} + \varrho(\frac{\partial^2 \mathcal{L}(\phi)}{\partial\phi\partial\phi'}|_{\phi=\phi^{l-1}})^{-1} gr(\phi^{l-1})$ where $\varrho > 0$ is a small scalar.

c) Modified Newton-Raphson.

The basic Newton-Raphson method requires the calculation of the matrix $\frac{\partial^2 \mathcal{L}(\phi)}{\partial\phi\partial\phi'}$ and its inversion. When ϕ is of large dimension this maybe computationally difficult. The modified Newton-Raphson method uses the fact that $\frac{\partial gr(\phi)}{\partial\phi} \approx \frac{\partial^2 \mathcal{L}(\phi)}{\partial\phi\partial\phi'}$ and guesses the shape of $\frac{\partial^2 \mathcal{L}(\phi)}{\partial\phi\partial\phi'}$ at the existing estimate using the derivative of the gradient. Let Σ^l be an estimate of $[\frac{\partial^2 \mathcal{L}(\phi)}{\partial\phi\partial\phi'}]^{-1}$ at iteration l. Then the method updates estimates of ϕ using (6.24) where

$$(\Sigma^l) = (\Sigma^{l-1}) + \frac{(-\rho^l \Sigma^{l-1} gr^{l-1})(-\rho^l \Sigma^{l-1} gr^{l-1})'}{(-\rho^l \Sigma^{l-1} gr^{l-1})\Delta gr^l} - \frac{\Sigma^{l-1}\Delta gr^l (\Delta gr^l)'(\Sigma^{l-1})^{-1}}{(\Delta gr^l)'\Sigma^{l-1}\Delta gr^l}$$

and $\Delta\phi^l = \phi^l - \phi^{l-1}$, $\Delta gr(\phi^l) = gr(\phi^l) - gr(\phi^{l-1})$. If likelihood is quadratic and the number of iterations large, $\lim_{l\to\infty} \phi^l = \phi_{ML}$ and $\lim_{l\to\infty} \Sigma^l = (\frac{\partial^2 \mathcal{L}(\phi)}{\partial\phi\partial\phi'}|_{\phi=\phi_{ML}})^{-1}$. Standard errors of the estimate can be read off the diagonal elements of Σ^l evaluated at ϕ_{ML}.

d) Scoring Method.

This method uses the information matrix $E\frac{\partial^2 \mathcal{L}(\phi)}{\partial\phi\partial\phi'}$ in place of $\frac{\partial^2 \mathcal{L}(\phi)}{\partial\phi\partial\phi'}$ in the calculation where the expectation is evaluated at $\phi = \phi^{l-1}$. The information matrix approximation is convenient since it typically has a simpler expression than the Hessian.

e) Gauss-Newton-scoring method.

The Gauss-Newton method uses a function of $(\frac{\partial e}{\partial\phi}|_{\phi=\phi^l})'(\frac{\partial e}{\partial\phi}|_{\phi=\phi^l})$ as an approximation to $\frac{\partial^2 \mathcal{L}(\phi)}{\partial\phi\partial\phi'}$, where ϕ_0^l is the value of ϕ at iteration l and e_t is the vector of errors in the model. In the case of constant state space models, the approximation is proportional to the vector of regressors constructed using the right hand side variables of both the state and the measurement equations. When the model is linear, Gauss-Newton and scoring approximations are identical.

6.4 ML estimation of DSGE models

Maximum likelihood estimation of the parameters of a DSGE model is a straightforward application of the methods we have described so far. As we have already seen in chapter 2, the log linearized solution of a DSGE model is of the form:

$$y_{2t} = \mathcal{A}_{22}(\theta)y_{2t-1} + \mathcal{A}_{23}(\theta)y_{3t} \quad (6.25)$$

$$y_{1t} = \mathcal{A}_{12}(\theta)y_{2t-1} + \mathcal{A}_{13}(\theta)y_{3t} \quad (6.26)$$

where y_{2t} includes the states and the driving forces, y_{1t} all other endogenous variables and y_{3t} the shocks of the model. Here $\mathcal{A}_{ii'}(\theta), i, i' = 1, 2$ are time invariant (reduced form) matrices which depend on $\theta = (\theta_1, \ldots, \theta_k)$, the structural parameters of preferences, technologies and government policies. Note also that there are cross equation restrictions in the sense that some $\theta_j, j = 1, \ldots, k$ may appear in more than one entry of these matrices.

Example 6.15 *In the working capital model considered in exercise 1.14 of chapter 2, setting $K_t = 1, \forall t$, y_{2t} includes lagged real balances $\frac{M_{t-1}}{p_{t-1}}$ and lagged deposits dep_{t-1}; y_{3t} includes shocks to the technology ζ_t and to the monetary rule M_t^g, while y_{1t} includes all the remaining endogenous variables (hours (n_t), output (GDP_t), the nominal interest rate (i_t) and the inflation rate (π_t)). Setting $N^{ss} = 0.33$, $\eta = 0.65$, $\pi^{ss} = 1.005$, $\beta = 0.99$, $(\frac{c}{GDP})^{ss} = 0.8$, the persistence of the shocks to 0.95 and the parameters of the policy rule to $a_2 = -1.0; a_1 = 0.5; a_3 = 0.1, a_0 = 0$, the log-linearizing solution has the following state space representation:*

$$\begin{bmatrix} \widehat{\frac{M_t}{p_t}} \\ \widehat{dep_t} \\ \widehat{GDP_t} \\ \widehat{n}_t \\ \widehat{i}_t \\ \widehat{\Pi}_t \end{bmatrix} = \begin{bmatrix} -0.4960 & 0.3990 \\ -1.0039 & 0.8075 \\ -0.3968 & 0.3192 \\ 0.9713 & -0.7813 \\ 2.0219 & -1.6264 \end{bmatrix} \begin{bmatrix} \widehat{\frac{M_{t-1}}{p_{t-1}}} \\ \widehat{dep_{t-1}} \end{bmatrix} + \begin{bmatrix} 1.3034 & -0.1941 \\ 1.1459 & -1.4786 \\ 1.0427 & -0.1552 \\ -0.3545 & 0.3800 \\ -0.9175 & -1.2089 \end{bmatrix} \begin{bmatrix} \widehat{\zeta_t} \\ \widehat{M^g}_t \end{bmatrix}$$

While in example 6.15 we have chosen a log-linear approximation, DSGE models with quadratic preferences and linear constraints also fit into this structure (see e.g. Hansen and Sargent (1998)). In fact, (6.25)-(6.26) are very general, do not require any certainty equivalence principle to obtain and need not be the solution to the model, as the next example shows.

Example 6.16 *(Watson) Suppose a model delivers the condition $E_t y_{t+1} = \alpha y_t + x_t$ where $x_t = \rho x_{t-1} + e_t^x$, x_0 given. This could be, e.g., a New-Keynesian Phillips curve, in which case x_t are marginal costs, or stock price relationship, in which case x_t are dividends. Using the innovation representation we have $x_t = E_{t-1}x_t + e_t^x$, $y_t = E_{t-1}y_t + e_t^y$ where $E_t x_{t+1} = \rho x_t = \rho(E_{t-1}x_t + e_t^x)$ and $E_t y_{t+1} = \alpha y_t + x_t = \alpha(E_{t-1}y_t + e_t^y) + (E_{t-1}x_t + e_t^x)$. Letting $y_{1t} = [x_t, y_t]$, $y_{2t} = [E_t x_{t+1}, E_t y_{t+1}]$, $y_{3t} = [e_t^x, v_t]$, where $v_t = e_t^y - E(e_t^y | e_t^x) = e_t^y - \kappa e_t^x$, $\mathcal{A}_{11}(\theta) = I$, $\mathcal{A}_{12}(\theta) = \begin{bmatrix} 1 & 0 \\ \kappa & 1 \end{bmatrix}, \mathcal{A}_{22}(\theta) = \begin{bmatrix} \rho & 0 \\ 1 & \alpha \end{bmatrix}, \mathcal{A}_{21}(\theta) = \begin{bmatrix} \rho & 0 \\ 1+\alpha\kappa & \alpha \end{bmatrix}$,*

it is immediate to see that the model fits into (6.25)-(6.26). Here the parameters to be estimated are $\theta = (\alpha, \rho, \kappa, \sigma_e^2, \sigma_v^2)$.

In general, one has two alternatives to derive a representation which fits (6.25)-(6.26): solve the model, as we have done in example 6.15, or use the rational expectations assumption, as we have done in example 6.16,

Exercise 6.19 *Consider a version of a consumption-saving problem where consumers are endowed with utility of the form* $u(c) = \frac{c^{1-\varphi}}{1-\varphi}$, *the economy is small relative to the rest of world and the resource constraint is* $c_t + B_t \leq GDP_t + (1 + r_t)B_{t-1}$ *where* B_t *are internationally traded bonds and* r_t *is the net real interest rate, taken as given by the agents.*
(i) Derive a log linearized version of the Euler equation and show how to map it into the framework described in example 6.16.
(ii) Show the entries of the matrices in the state space representation.
(iii) How would you include a borrowing constraint $B_t < \bar{B}$ *in the setup?*

Exercise 6.20 *Consider the labor hoarding model studied in exercise 4.1 of chapter 5 where agents have preferences over consumption, leisure and effort and firms distinguish between labor and effort in the production function. Cast the log-linearized Euler conditions into a state space framework using an innovation representation.*

Clearly, (6.25)-(6.26) are in a format estimable with the Kalman filter. In fact, recursive estimates of y_{2t} can be obtained, given some initial conditions y_{20}, if $\mathcal{A}_{ii'}(\theta)$, $\sigma_{y_3}^2$ are known. Given these recursive estimates forecast errors can be computed. Hence, for each choice of θ, we can calculate the likelihood function via the prediction error decomposition and update estimates using one of the algorithms described in section 3. Standard errors for the estimated parameters can be read off the Hessian, evaluated at maximum likelihood estimates, or any approximation to it.

Despite the simplicity of this procedure, there are several issues, specific to DSGE models, one must deal with when using ML to estimate structural parameters. The first has to do with the number of series used in the estimation. As it is clear from (6.25)-(6.26), the covariance matrix of the vector $[y_{1t}, y_{2t}]$ is singular, a restriction unlikely to hold in the data. This singularity was also present in the innovation representation (6.11). Two options are available to the applied investigator: she can either select as many variables as there are shocks or artificially augment the space of shocks with measurement errors. For example, if the model is driven by a technology and a government expenditure shock, one selects two of (the many) series belonging to $[y_{1t}, y_{2t}]$ to estimate parameters. Kim (2000) and Ireland (2000) use such an approach in estimating versions of sticky price models. While this leaves some arbitrariness in the procedure, some variables may have little information about the parameters. Although a-priori it may hard to know which equations carry information, one could try to select variables so has to maximize the identifiability of the parameters. Alternatively, since some variables may not satisfy the assumptions needed to obtain consistent estimates (for example, they display structural breaks), one could choose the variables that are more likely to satisfy these assumptions.

Example 6.17 *In a standard RBC model driven by technology disturbances we have that* $[\hat{N}_t, \widehat{gdp}_t, \hat{c}_t]$ *are statically related to the states* $[\hat{K}_t, \hat{\zeta}_t]$ *via a matrix* $\mathcal{A}_{12}(\theta)$ *where* $\hat{\cdot}$ *refers to deviations from the steady state. Since the number of shocks is less than the number of endogenous variables, there are linear combinations of the controls which are perfectly predictable. For example, using equation (6.25) into (6.26) we have that* $\alpha_1 \hat{N}_t + \alpha_2 \widehat{gdp}_t + \alpha_3 \hat{c}_t = 0$ *where* $\alpha_1 = \mathcal{A}_1^{11}\mathcal{A}_1^{32} - \mathcal{A}_1^{12}\mathcal{A}_1^{31}$, $\alpha_2 = \mathcal{A}_1^{22}\mathcal{A}_1^{31} - \mathcal{A}_1^{21}\mathcal{A}_1^{32}$, $\alpha_3 = \mathcal{A}_1^{22}\mathcal{A}_1^{11} - \mathcal{A}_1^{31}\mathcal{A}_1^{21}$. *Similarly, using the equations for* $\widehat{gdp}_t, \hat{c}_t$ *and the law of motion of the capital stock we have* $\alpha_4 \hat{c}_t + \alpha_5 \hat{c}_{t-1} - \alpha_6 \widehat{gdp}_t - \alpha_7 \widehat{gdp}_{t-1} = 0$ *where* $\alpha_4 = \mathcal{A}_1^{12} + \delta[1 - \delta(\frac{K}{N})^{\eta}](\mathcal{A}_1^{12}\mathcal{A}_1^{31} - \mathcal{A}_1^{11}\mathcal{A}_1^{32}/[1 - \delta(\frac{K}{N})^{\eta}]$, $\alpha_5 = (1-\delta)\mathcal{A}_1^{12}$ $\alpha_6 = \mathcal{A}_1^{32} - \delta(\mathcal{A}_1^{12}\mathcal{A}_1^{31} - \mathcal{A}_1^{12}\mathcal{A}_1^{32})/[1 - \delta(\frac{K}{N})^{\eta}]$, $\alpha_7 = (1-\delta)\mathcal{A}_1^{32}$. *This implies that the system is stochastically singular and for any sample size the covariance matrix of the data is postulated to be of reduced rank.*

Attaching measurement errors to (6.26) is the option taken by Sargent (1979), Altug (1989) or McGrattan, Rogerson and Wright (1997). The logic is straightforward: by adding a vector of serially and contemporaneously uncorrelated measurement errors, we complete the probability space of the model (the theoretical covariance matrix of $[y_{1t}, y_{2t}]$ is no longer singular). Since actual variables typically fail to match their model counterparts (e.g. actual savings are typically different from model based measures of savings), the addition of measurement errors is easily justifiable. Note that, if this route is taken, a simple diagnostic on the quality of the model can be obtained by comparing the size of the estimated standard deviation of the measurement errors and of the structural shocks. Standard deviations for the former much larger than for the latters suggest that misspecification is likely to be present.

Example 6.18 *In example 6.15, if we wish to complete the probability space of the model, we need to add five measurement errors to the vector of shocks. Alternatively, we could use, e.g., real balances and deposits to estimate the parameters of the model. However, it is unlikely that these two series have information to estimate the share of labor in production function* η. *Hence, identification of the parameters may be a problem when using a subset of the variables of the model*

The introduction of a vector of serially and contemporaneously uncorrelated measurement errors does not alter the dynamics of the model. Therefore, the quality of the model's approximation to the data is left unchanged. Ireland (2004), guessing that both dynamic and contemporaneous misspecifications are likely to be present in simple DSGE models, instead adds a VAR(1) vector of measurement errors. The importance of these dynamics for the resulting hybrid model can be used to gauge how far the model is from the data, much in the spirit of Watson (1993), and an analysis of the properties of the estimated VAR may help in respecifying the model (see chapter 7). However, is important to note that the hybrid model can no longer be considered "structural": the additional dynamics play the same role as distributed lags which were added in the past to specifications derived from static economic theory when confronted with the (dynamics of the) data.

The second issue concerns the quality of the model's approximation to the data. It is clear that to estimate the parameters with ML and to validate model, one must assume that it "correctly" represents the process generating the data up to a set of unknown parameters. Some form of misspecification regarding e.g. the distribution of the errors (see White (1982)) or the parametrization (see Hansen and Sargent (1998)), can be handled using the quasi-maximum likelihood approach discussed in section 6.2. However, as we will argue in chapter 7, the misspecification that a DSGE model typically displays is of different type. Adding contemporaneous uncorrelated measurement errors avoids singularities but it does not necessarily reduce misspecification. Moreover, while with GMM one is free to choose the relationships used to estimate the parameters of interest, this is not the case with ML since joint estimation of all the relationships produced by the model is generally performed. Under these conditions, maximum likelihood estimates of the parameters are unlikely to be consistent and economic exercises conducted conditional on these estimates may be meaningless. In other words, credible maximum likelihood estimation of the parameters of a DSGE model requires strong beliefs about the nature of the model.

Third, for parameters to be estimable they need to be identifiable. For example, if θ_1 and θ_2 are parameters and only $\theta_1 + \theta_2$ or $\theta_1\theta_2$ are identifiable, they can not be estimated separately. Besides this generic problem, thoroughly discussed in Hamilton (1994), DSGE models often face partial identifiability problems in the sense that the series used may have little information about the parameters of interest. This is not surprising: estimating, say, parameters of a monetary policy rule out of export or the trade balance is unlikely to be successful even if these parameters appear in the relevant equations. Furthermore, certain parameters affect only the steady state and therefore cannot be estimated when the model is written in deviations from the steady states or when variables are entered in log differences. In this situation two approaches are possible. The first one, which is more standard, is to calibrate nonestimable parameters (say θ_1) and provide ML estimates for the remaining free parameters (say θ_2) conditional on the chosen θ_1. As argued in chapter 7, such a choice may generate consistency problems and distort the asymptotic distribution of θ_2. The alternative is to use other moment conditions were these parameters appear and jointly estimate θ_1 and θ_2 using the scores of the likelihood and these moment conditions. Since the score of the likelihood has the format of moment conditions, this mixed approach will produce, under regularity conditions, consistent and asymptotically normal estimates. When this last alternative is unfeasible, local sensitivity analysis in a neighborhood of the calibrated parameters is advisable to explore the shape of the likelihood function around the maximum for θ_2 one finds.

Note also the similarities between this and the GMM approach described in chapter 5. Two main differences should be noted. First, the construction of the scores requires the solution of the model (or the rational expectation assumption), which was not necessary to estimate parameters with GMM. Second, if no misspecification is present, ML estimates will, by construction, be more efficient than GMM estimates.

Once parameter estimates are obtained one can proceed to validate the model and/or examine the properties of the implied system. Statistical validation can be conducted in

many ways. For example, if interesting economic hypotheses involve restrictions on a subset of the parameters of the model, standard t-tests or likelihood ratio tests using the restricted and the unrestricted versions of the model can be performed.

Example 6.19 *(Money demand equation) Consider a representative agent maximizing* $E_0 \sum_t \beta^t \left[\frac{1}{1-\varphi_c} c_t^{1-\varphi_c} + \frac{\vartheta_M}{1-\varphi_M}\left(\frac{M_{t+1}}{p_t}\right)^{1-\varphi_M}\right]$ *by choice of* (c_t, B_{t+1}, M_{t+1}) *subject to* $c_t + \frac{B_{t+1}}{p_t} + \frac{M_{t+1}}{p_t} + \frac{b_1}{2}\frac{(M_{t+1}-M_t)^2}{p_t} + \frac{b_2}{2}\frac{(M_t - M_{t-1})^2}{p_t} \leq w_t + \frac{M_t}{p_t} + (1+i_t)\frac{B_t}{p_t}$*, where* b_1, b_2 *are parameters,* w_t *is an exogenous labor income and* B_t *are nominal one-period bonds. The two optimality conditions are* $c_t^{-\varphi} = \beta E_t[c_{t+1}^{-\varphi}\frac{p_t}{p_{t+1}}(1+i_{t+1})]$ *and* $\vartheta_M\left(\frac{M_{t+1}}{p_t}\right)^{-\varphi_M} c_t^{\varphi_c} = E_t[1 - \frac{1}{1+i_{t+1}} + (b_1 + \frac{b_2}{1+i_{t+1}})\Delta M_{t+1} - \frac{1}{1+i_{t+1}}(b_1 + \frac{b_2}{1+i_{t+2}})\Delta M_{t+2}]$ *where* $\Delta M_{t+1} = M_{t+1} - M_t$*. Log linearizing the two conditions, solving out for* $\hat{i}_{t+1}$ *and using the budget constraint we have that* $\phi_c \hat{w}_t - \phi_M(\hat{M}_{t+1} - \hat{p}_t) = \alpha_1 \widehat{\Delta M}_{t+1} + \alpha_2 \widehat{\Delta M}_{t+2} + \alpha_3 \widehat{\Delta w}_{t+1} + \alpha_4 \widehat{\Delta w}_{t+2} + \alpha_5 \widehat{\Delta p}_{t+1} + \alpha_6 \widehat{\Delta p}_{t+2}$ *where* α_j *are functions of the deep parameters of the model* (b_1, b_2) *and of the steady states* $i^{ss}, \Delta M^{ss}$*. If we assume that the Central bank chooses* i_{t+1} *so that* $\Delta \hat{p}_t = 0$*, that bonds are in zero net supply, the above equation can be solved for* ΔM_t *as a function of the current and future exogenous labor income* $\hat{w}_t$ *and the current and future levels of real balances* $\hat{M}_{t+1} - \hat{p}_t$*.*

The parameters of this model can be estimated in a number of ways. One is GMM. For example, using as instruments lagged values of money growth, real balances and labor income, one could estimate $(\varphi_M, \varphi_c, b_1, b_2, i^s s, \Delta M^{ss}, \beta)$ *from the above equation. Alternatively, one could use ML. To do so the above equation needs to solved forward in order to express current growth rate of money as a function of current and future consumption and current money holdings. As we will see in example 6.20, this is easier to do if we represent the available data with a VAR.*

Since there is only one shock (the exogenous labor income) and the system of equations determining the solution is singular. There are three alternatives to deal with this problem. The one we have used expresses the solution of ΔM_t *in terms of current and future labor income and real balances. Then estimates of the parameters can be found maximizing the likelihood of the resulting equation. The second is to attach to the policy equation an error,* $\Delta \hat{p}_t = \epsilon_{3t}$*. This is easily justifiable if inflation targeting is only pursued on average over some period of time. The third is to assume that labor income is measured with error. In the latter two alternatives, the joint likelihood function of the money demand equation and of the consumption Euler equation can be used to find estimates of the parameters. Note also that not all the parameters may be identifiable from the first setup - the forward looking solution requires elimination of the unstable roots which may have important information about, e.g., the adjustment cost parameters.*

Restricted and unrestricted specifications can also be compared in an out-of-sample forecasting race; for example, using the MSE of the forecasts, or the record of turning point predictions.

Exercise 6.21 *Consider two versions of a RBC model, one with capacity utilization and one without. Describe a Monte Carlo procedure to verify which model matches turning*

points of US output growth better. How would you compare models which are not nested (say, one with capacity utilization and one with adjustment costs to capital)?

The stability of the estimates over subsamples can be examined in a standard way. For example, one can split the sample in two and construct a distance test of the form $\mathsf{S} = (\theta^1 - \theta^2)(\Sigma_{\theta^1} + \Sigma_{\theta^2})^{-1}(\theta^1 - \theta^2)$ where θ^1 is the ML estimate obtained in the first sample and Σ_{θ^1} its estimated covariance matrix and θ^2 is the ML estimate obtained in the second sample and Σ_{θ^2} the corresponding estimated covariance matrix. Recursive tests of this type can also be used to determine when a structural break occurs. That is, for each $1 < \tau < T$, we can construct S_τ by estimating the model over two samples $[1, \tau], [\tau + 1, T]$. Then one would compare $\sup_\tau \mathsf{S}_\tau$ to a $\chi^2(\dim(\theta))$, much in the same spirit as structural stability tests described in chapter 4.

We have seen that the solution of DSGE models can be alternatively written in a state space or restricted VAR(1) format. This latter offers an alternative framework to compare the model to the data. The restrictions that DSGE imposes on VARs are of two types. First, log-linearized DSGE models are typically VAR(1) models. Therefore, the methods described in chapter 4 can be used to examine whether the actual data can be modelled as an VAR(1). Second, it is well known, at least since Sargent (1979), that rational expectations models impose an extensive set of cross equations restrictions on the VAR of the data. These restrictions can be used to identify and estimate the free parameters and to test the validity of the model. We discuss how this can be done next.

Example 6.20 *(Kurmann) Consider an hybrid Phillips curve, $\pi_t = \alpha_1 E_t \pi_{t+1} + \alpha_2 \pi_{t-1} + \alpha_3 mc_t + e_t$ which can be obtained from a standard sticky price model once a fraction of the producers fix the price using a rule of thumb and adding some measurement error e_t. The rule necessary to produce such an expression is that the new price is set to an average of last period's price, updated with last period's inflation rate (as in Gali and Gertler (1999)). Assume mc_t is exogenous and let $\mathcal{F}_t$ represents the information set available at each t. For any $z_t \in \mathcal{F}_t$, $E_t(E_t[y_{t+\tau}|\mathcal{F}_t]|z_t) = E_t(y_{t+\tau}|z_t)$, by the law of iterated expectations. Let $\mathbb{Y}_t = \mathbb{A}\mathbb{Y}_{t-1} + E_t$ be the companion form representation of the model where $\mathbb{Y}_t$ is of dimension $mq \times 1$ (m variables with q lags each). Since $E_t(mc_{t+\tau}|\mathbb{Y}_t) = \mathcal{S}_1 \mathbb{A}^\tau \mathbb{Y}_t$ and $E(\pi_{t+\tau}|\mathbb{Y}_t) = \mathcal{S}_2 \mathbb{A}^\tau \mathbb{Y}_t$ where $\mathcal{S}_1$ and $\mathcal{S}_2$ are selection matrices, a hybrid Phillips curve implies $\mathcal{S}_2[\mathbb{A} - \alpha_1 \mathbb{A}^2 - \alpha_2 I] = \alpha_3 \mathcal{S}_1 \mathbb{A}$ which produce mq restrictions. For example, if $q = 1$, $\mathbb{Y}_t$ includes only the labor share and inflation and $A_{ii'}$ are the parameters of the VAR we have*

$$\begin{aligned} A_{12} - \alpha_1 A_{12} A_{11} - \alpha_1 A_{22} A_{12} - \alpha_2 &= \alpha_3 A_{11} \\ A_{22} - \alpha_1 A_{21} A_{12} - \alpha_1 A_{22}^2 - \alpha_2 &= \alpha_3 A_{21} \end{aligned} \tag{6.27}$$

(6.27) requires that expectations of real marginal costs and inflation produced by a VAR are consistent with the dynamics of the model. One way to impose these restrictions is to express the coefficients of the inflation equation in the VAR as a function of the remaining $(m-1)mq$ VAR coefficients and the parameters of the theory. Here, since there are four

unknowns and two equations, the system can be solved for, e.g., A_{21} and A_{22} as a function of A_{11} and A_{12}. The likelihood function for the restricted VAR system can then be constructed using the prediction error decomposition and tests of the restrictions obtained comparing the likelihood of restricted and unrestricted VARs.

Exercise 6.22 *Consider an endowment economy where agents receive a random income y_t and may either consume or save it. Suppose that stocks S_{t+1} are the only asset, that their price is p_t^s and that the budget constraint is $c_t + p_t^s S_{t+1} = y_t + (p_t^s + sd_t)S_t$ where sd_t are dividends. Assume $u(c) = \frac{c^{1-\varphi}}{1-\varphi}$ and that agents discount the future at the rate β.*
i) Derive a log-linearized expression for the price of stocks as a function of future dividends, future prices and current and future consumption.
ii) Assume that data on stock prices and stock dividends are available and that an econometrician specifies the process for the data as a VAR of order 2. Derive the cross-equation restrictions that the model imposes on the bivariate representation of prices and dividends (Hint: use the equilibrium conditions to express consumption as a function of dividends).
iii) Assume that also data on consumption is available. Does your answer in ii) change?

Exercise 6.23 *Continuing with example 6.19, consider the log-linearized money demand equation alone. Assume that $\Delta X_t = [\Delta p_t, \Delta w_t, \Delta i_t]$ is exogenous and follows a VAR(q) model which we write in the companion form $\mathbb{Y}_t = \mathbb{A}\mathbb{Y}_{t-1} + \mathbb{E}_t$ and $\Delta X_t = \mathbb{S}\mathbb{Y}_t$, where $\mathbb{S}$ is a selection matrix. Show that the forward solution can be written as $\Delta \ln M_t = \frac{M_0}{1-\lambda_1} - (1 - \frac{1+i^{ss}}{\lambda_1}(\ln M_{t-1} - \tilde{\phi}'\mathbb{S}\mathbb{Y}_{t-1}) + \frac{\lambda_1 - 1 - (1+i^{ss})}{\lambda_1 - 1}\tilde{\phi}\mathbb{S}(1 - \frac{\mathbb{A}}{\lambda_1})^{-1}\mathbb{Y}_t + v_t$ where $\phi = [1, \frac{\phi_c}{\phi_M}, -\frac{1}{i^s s \phi_M}]$, $\tilde{\phi} = \frac{i^{ss}\varphi_M}{M^{ss}(b_1 + b_2/(1+i^{ss}))(\lambda_1 - (1+i^{ss}-1)}\phi$, λ_1 is the stable solution of $\lambda^2 - (1 + (1+i^{ss}) + (\frac{i^{ss}\varphi_M}{M(b_1 + b_2/(1+i^s s)}) \lambda + 1 = 0$, and v_t is a measurement error, which appears because the econometrician information may be different from the one of the agents. Give the structure of v_t. Show the format of the solution when $q = 2$ and there is no constant in the VAR for ΔX_t. What parameters can you estimate? Write down the likelihood function you want to maximize and show the implied cross equations restrictions.*

Statistical validation is usually insufficient for economic purposes, since it offers scarce indications on the reasons why the model fails and provides very little information about the properties of the estimated model. Therefore, as we have done in chapter 5, one would also like to compare the predictions of the model for a set of interesting statistics of the data. Several statistics can be used. For example, given ML estimates, one could compute unconditional moments such as variability, cross correlations, spectra or cross spectra and compared them with those in the data. To learn about the dynamic properties of the estimated model one could compute impulse responses, variance and historical decompositions. Informal comparisons are typically performed but there is no reason to do so, especially in a ML context. In fact, since $\sqrt{T}(\theta_{ML} - \theta_0) \xrightarrow{D} \mathbb{N}(0, \Sigma_\theta)$, we can compute the asymptotic distribution of any continuous function of θ using the δ-method i.e. if $h(\theta)$ is continuously differentiable, $\sqrt{T}(h(\theta_{ML}) - h(\theta_0)) \xrightarrow{D} N(0, \Sigma_h = \frac{\partial h(\theta)}{\partial \theta}\Sigma_\theta \frac{\partial h(\theta)}{\partial \theta}')$. If an estimate h_T is available in the data, a formal measure of distance between the model and the

data is $(h(\theta_{ML}) - h_T)(\Sigma_h + \Sigma_{h_T})^{-1}(h(\theta_{ML}) - h_T)$, which is asymptotically distributed as a $\chi^2(dim(h))$. Small sample versions of such tests are also easily designed.

Exercise 6.24 *Suppose that* $\sqrt{T}(\theta_{ML} - \theta_0) \xrightarrow{D} N(0, \Sigma_\theta)$ *and suppose that, for the statistic* $h(\theta)$ *of interest, both* h_T *and its standard error are available. Describe how to perform a small sample test of the fit of the model.*

Once the model is found to be adequate in capturing the statistical and the economic features of the data, welfare measures can be calculated and policy exercises performed.

Exercise 6.25 *(Blanchard and Quah) The model described in section 3.1 of chapter 3 produces a solution of the form*

$$\begin{aligned} \Delta GDP_t &= \epsilon_{3t} - \epsilon_{3t-1} + (1+a)\epsilon_{1t} - a\epsilon_{1t-1} \\ UN_t &= -\epsilon_{3t} - a\epsilon_{1t} \end{aligned} \tag{6.28}$$

where $\Delta GDP_t = GDP_t - GDP_{t-1}$ *and* $UN_t = N_t - N^{fe}$ *where* N^{fe} *is full employment equilibrium,* ϵ_{1t} *is a technology shock and* ϵ_{3t} *a money shock.*
i) Transform (6.28) into a state space model
ii) Using data for output growth and appropriately detrended unemployment provide a maximum likelihood estimate of α *and test three hypotheses,* $\alpha = 0$ *and* $\alpha \pm 1$.
iii) Provide impulse responses to technology and money shocks using α_{ML}. *Compare them with those obtained with a structural VAR identified using long run restrictions.*

Exercise 6.26 *(Habit persistence) Consider a basic RBC model driven by technology disturbances and three separate specifications for preferences. The first one assumes intertemporal separability of consumption and leisure, that is* $u(c_t, c_{t-1}, N_t, N_{t-1}) = \frac{c_t^{1-\varphi}}{1-\varphi} + \ln(1-N_t)$. *The second that there is habit persistence in consumption, that is* $u(c_t, c_{t-1}, N_t, N_{t-1}) = \frac{(c_t+\gamma c_{t-1})^{1-\varphi}}{1-\varphi} + \ln(1-N_t)$. *The third that there is habit persistence in leisure so that* $u(c_t, c_{t-1}, N_t, N_{t-1}) = \frac{c_t^{1-\varphi}}{1-\varphi} + \ln(1-N_t+\gamma(1-N_{t-1}))$. *The resource constraint is* $c_t + K_{t+1} = \zeta_t K_t^{1-\eta} N_t^{\eta} + (1-\delta)K_t$ *where* $\ln \zeta_t$ *is an AR(1) process. Using US data on consumption, hours, output and investment estimate the free parameters of the three models assuming that consumption, investment and output are measured with error and that each of these errors is a contemporaneously uncorrelated martingale difference process. Test the hypotheses that habit persistence either in consumption or in leisure is unnecessary to match the data. Compare the responses of the three models to technology shocks. What is the role of habit persistence in propagating technology disturbances? (Hint: Nest the three models in one general specification and test the restrictions).*

Exercise 6.27 *(Woodford) Suppose agents maximize* $E_0 \sum_t \beta^t \epsilon_{4t}[u_1(c_t + G_t) + u_2(\frac{M_t}{p_t}) - \epsilon_{2t} u_3(N_t)]$ *where* G_t *is government expenditure,* $\frac{M_t}{p_t}$ *are real balances,* N_t *is hours,* $c_t = (\int c_{it}^{\frac{1}{\varsigma_p+1}} di)^{\varsigma_p+1}$ *and* $p_t = (\int p_{it}^{-\frac{1}{\varsigma_p}} dj)^{-\varsigma_p}$. *Here* ϵ_{4t} *is a aggregate demand shock and* ϵ_{2t}

a labor supply shock and ς_p the elasticity of substitution across consumption goods. Let aggregate demand for good i be $c_{it} = c_t(\frac{p_{it}}{p_t})^{-\frac{1+\varsigma_p}{\varsigma_p}}$. The budget constraint of consumers is $c_t + \frac{M_t}{p_t} + \frac{B_t}{p_t} + T_t = w_t N_t + \frac{M_{t-1}}{p_t} + \frac{(1+i_{t-1})}{\pi_t}\frac{B_{t-1}}{p_{t-1}}$, where $\frac{B_t}{p_t}$ are real bonds and π_t the inflation rate. Suppose $c_{it} = N_{it}$ and that the price index evolves according to $p_t = (\zeta_p p_{t-1}^{-\frac{1}{\varsigma_p}} + (1-\zeta_p)\tilde{p}_t^{-\frac{1}{\varsigma_p}})^{-\varsigma_p}$, where $\tilde{p}_t$ is the optimal price in a Calvo style setting and ζ_p the fraction of firms not changing prices. Finally, assume that the monetary authority sets interest rates according to $1 + i_t = a_1 + a_2\pi_t + (1+i^{ss})M_t^g$, where i^{ss} is the steady state net interest rate and the fiscal authorities sets T_t according to $T_t = a_3 + a_4\frac{B_{t-1}}{p_{t-1}} + T^{ss}T_t^g$ where T^{ss} are steady state lump sum taxes.
i) Derive the log linearized first order conditions (around the steady states) of the model.
ii) Derive a state space representation for the conditions in i) in terms of $\hat{\epsilon}_t = [\hat{\epsilon}_{4t}, \hat{\epsilon}_{2t}, \hat{M}_t^g, \hat{T}_t^g]$.
iii) Assuming that $\hat{\epsilon}_t$ is an AR(1) with diagonal persistence matrix and that output and inflation are measured with error, provide ML estimates of the parameters of the model using US data for debt, real balances, inflation, output, nominal interest rate and real deficit. Test the hypothesis $a_4 < \frac{1-\beta}{\beta}$ and $a_2 < \frac{1}{\beta}$, which corresponds to passive fiscal policy and active monetary policy in the terminology of Leeper (1991). What is the effect of shocks to T_t^g in the economy?

6.5 Estimating a sticky price model: an example

The model we consider is the same as in exercise 3.2 of chapter 3. Our task is to estimate its structural parameters, test interesting economic hypotheses concerning the magnitude of the coefficients, compare the forecasting performance relative to an unrestricted VAR and, finally, compare some conditional moment implications of the model and of the data.

For convenience we repeat the basic setup: the representative household maximizes $E_0\sum_t \beta^t[\ln c_t + \vartheta_M \ln(\frac{M_t}{p_t}) - \frac{\vartheta_N}{1-\varphi_n}N_t^{1-\varphi_n} - \frac{\vartheta_{ef}}{1-\varphi_{ef}}Ef_t^{1-\varphi_{ef}}]$ where $c_t = (\int c_{it}^{\frac{1}{\varsigma_p+1}}di)^{\varsigma_p+1}$ is aggregate consumption, ς_p is the elasticity of substitution among consumption goods, $p_t = (\int p_{it}^{-\frac{1}{\varsigma_p}}dj)^{-\varsigma_p}$ is the aggregate price index, $\frac{M_t}{p_t}$ are real balances, N_t is hours worked and Ef_t is effort. The budget constraint is $\int_0^1 p_{it}c_{it}di + M_t = W_{Nt}N_t + W_{et}Ef_t + M_{t-1} + T_t + Prf_t$ where T_t are monetary transfers, Prf_t profits distributed by the firms and W_{Nt}, W_{et} are the reward to working and to effort. A continuum of firms produce differentiated good using $c_{it} = \zeta_t(N_{it}^{\eta_2}Ef_{it}^{1-\eta_2})^{\eta_1}$ where $N_{it}^{\eta_2}Ef_{it}^{1-\eta_2}$ is the quantity of effective input and ζ_t an aggregate non-stationary technology shock, $\Delta\zeta_t = \epsilon_{1t}$ where $\ln\epsilon_{1t} \sim iid\ \mathbb{N}(0, \sigma_\zeta^2)$. Firms set prices one period in advance, taking as given the aggregate price level and not knowing the current realization of the shocks. Once shocks are realized, firms optimally choose employment and effort. So long as marginal costs are below the predetermined price, firms will meet the demand for their product and choose an output level equal to $c_{it} = (\frac{p_{it}}{p_t})^{-1-\varsigma_p^{-1}}c_t$. Optimal price setting implies $E_{t-1}[\frac{1}{c_t}((\eta_1\eta_2)p_{it}c_{it} - (\varsigma_p+1)W_{Nt}N_{it})] = 0$ which, in the absence of uncertainty, reduces to the standard that condition that the price

is a markup over marginal costs. We assume that the monetary authority controls the quantity of money and sets $\Delta M_t = \epsilon_{3t} + a_M \epsilon_{1t}$ where $\ln \epsilon_{3t} \sim iid\ \mathbb{N}\ (0, \sigma_M^2)$ and a_M is a parameter. Letting lower case letters denote natural logs, the model implies the following equilibrium conditions for inflation (Δp_t), output growth (Δgdp), employment (n_t) and labor productivity growth (Δnp_t)

$$\begin{aligned}
\Delta p_t &= \epsilon_{3t-1} - (1 - a_M)\epsilon_{1t-1} && (6.29)\\
\Delta gdp_t &= \Delta\epsilon_{3t} + a_M\epsilon_{1t} + (1 - a_M)\epsilon_{1t-1} && (6.30)\\
n_t &= \frac{1}{\eta}\epsilon_{3t} - \frac{1 - a_M}{\eta}\epsilon_{1t} && (6.31)\\
\Delta np_t &= (1 - \frac{1}{\eta})\Delta\epsilon_{3t} + (\frac{1 - a_M}{\eta} + a_M)\epsilon_{1t} + (1 - a_M)(1 - \frac{1}{\eta})\epsilon_{1t-1} && (6.32)
\end{aligned}$$

where $np_t = gdp_t - n_t$ and $\eta = \eta_1(\eta_2 + (1 - \eta_2)\frac{1+\varphi_n}{1+\varphi_{ef}})$.

The model therefore has two shocks (a technology and a monetary one) and implications for at least four variables ($\Delta p_t, \Delta gdp_t, \Delta np_t, n_t$). There are 11 free parameters ($\eta_1, \eta_2, \varphi_n, \varphi_{ef}, \beta, \sigma_\zeta^2, \sigma_M^2, a_M\ \vartheta_M, \vartheta_n, \vartheta_{ef}$), but many of them are do not appear in or are not identifiable from (6.29)-(6.32). In fact it is easy to verify that only a_M and η independently enter the four conditions and therefore, together with σ_ζ^2 and σ_M^2, are the only ones estimable with likelihood methods.

Since there are only two shocks the covariance matrix produced by the model is singular and we are free to choose which two variables to use to estimate the parameters. In the baseline case we select productivity and hours. As a robustness check, we repeat estimation using both output and hours, and prices and output. Note that, in this latter case, also η is non-identifiable. As an alternative, we estimate the model adding serially uncorrelated measurement errors to output and productivity. In this case we estimate six parameters: the four structural ones and the variances of the two measurement errors.

We examine both the statistical and economic fit of the model. First, we study several specifications which restrict a_M and/or η to some prespecified value. A Likelihood ratio test is performed in each case and the statistics compared to a χ^2 distribution. For the specification with measurement errors, we also perform a forecasting exercise comparing the one step ahead MSE of the model to the MSE produced by a four variable VAR(1) model, which has 20 parameters (four constants and 16 autoregressive coefficients). Since the number of coefficients in the two specifications differs, we also compare the two specifications with a Schwarz criterion (see chapter 4). In this latter case, the VAR model is penalized since it has a larger number of parameters. We also compute tests of forecasting accuracy, as detailed in section 6.2. Conditional on the estimated parameters, we compute impulse responses, to examine the sign of the dynamics of the variables to technology and monetary shocks, and compare few elements of the unconditional autocovariance function for the four variables in the model and in the data.

We use CPI, GDP (constant in 1992 prices) and total hours (equal to average weekly hours multiplied by civilian employment) for Canada for the period 1981:2-2002:3. All

variables are logged and first differences of the log are used to compute growth rates. Total hours are detrended using a linear trend.

(6.29)-(6.32) has a state space representation for $\alpha = [\epsilon_{1t}, \epsilon_{1t-1}, \epsilon_{3t}, \epsilon_{3t-1}, v_{1t}, v_{2t}]$, where $v_{it}, i = 1, 2$ are measurement errors, $x_{1t} = \begin{bmatrix} 0 & a_M - 1 & 0 & 1 & 0 & 0 \\ a_M & 1 - a_M & 1 & -1 & 1 & 0 \\ \frac{a_M - 1}{\eta} & 0 & \frac{1}{\eta} & 0 & 0 & 0 \\ \frac{1-a_M}{\eta} + a_M & \frac{(1-a_M)(\eta-1)}{\eta} & \frac{\eta-1}{\eta} & -\frac{\eta-1}{\eta} & 0 & 1 \end{bmatrix}$,

$\mathbb{D}_1 = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$, $\mathbb{D}_2 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$, $\Sigma_{v_1} = 0$ with the appropriate adjustments if no measurement error is included. The Kalman filter is initialized using $\alpha_{1|0} = 0$ and $\Omega_{1|0} = I$. The likelihood function is computed recursively and a simplex method is used to locate the maximum. We use this approach instead of a one based on the gradient because the likelihood is flat, the maximum is around the boundary of the parameter space and convergence is hard to achieve. The cost is that no standard errors for the estimates are available. Table 6.1 reports parameter estimates, together with the p-values of various likelihood ratio tests.

Table 6.1: ML estimates

Data set	a_M	η	σ^2_ζ	σ^2_M			Log Likelihood
$(\Delta np_t, n_t)$	0.5533	0.9998	1.06e-4	6.69e-4			704.00
$(\Delta gdp_t, n_t)$	-7.7336	0.7440	6.22e-6	1.05e-4			752.16
$(\Delta gdp_t, \Delta p_t)$	3.2007		1.26e-5	1.57e-4			847.12
	a_M	η	σ^2_ζ	σ^2_M	σ^2_{v1}	σ^2_{v2}	Log Likelihood
$(n_t, \Delta np_t, \Delta gdp_t, \Delta p_t)$	-0.9041	1.2423	5.82e-6	4.82e-6	0.0236	0.0072	1336
Restrictions	$a_M = 0$	$\eta = 1$	$\eta = 1$	$\eta = 1.2$			
	$a_M = -1.0$						
$(\Delta np_t, n_t)$, p-value	0.03	0.97	0.01	0.00			
$(\Delta gdp_t, n_t)$, p-value	0.00	0.00	0.00	0.00			
$(n_t, \Delta np_t, \Delta gdp_t, \Delta p_t)$ p-value	0.00	0.001	0.00	0.87			
Restrictions	$a_M = 0$	$a_M = 1$	$a_M = -1.0$				
$(\Delta y_t, \Delta p_t)$, p-value	0.00	0.00	0.00				

Several features of the table deserve comments. First, using bivariate specifications the estimated value of η is less then one. Since for $\varphi_{ef} = \varphi_N$, $\eta = \eta_1$, this implies that there is no evidence of short run increasing returns to scale. The lack of increasing returns is formally confirmed by likelihood ratio tests: conditioning on values of $\eta \geq 1$ reduces the likelihood. However, when measurement errors are included, mild short run increasing returns to scale obtain. Second, the estimated value of a_M depends on the data set: it is positive and moderate when productivity and hours are used; positive and large when output and prices

are used, strongly negative when output and hours are used and moderately negative when the four series are used. The reason for this large variety of estimates is that the likelihood function is very flat in the a_M dimension. Figure 6.1 illustrates this fact using the first data set. It is easy to see that $a_M = 0$, or $a_M = -0.5$ are not extremely unlikely.

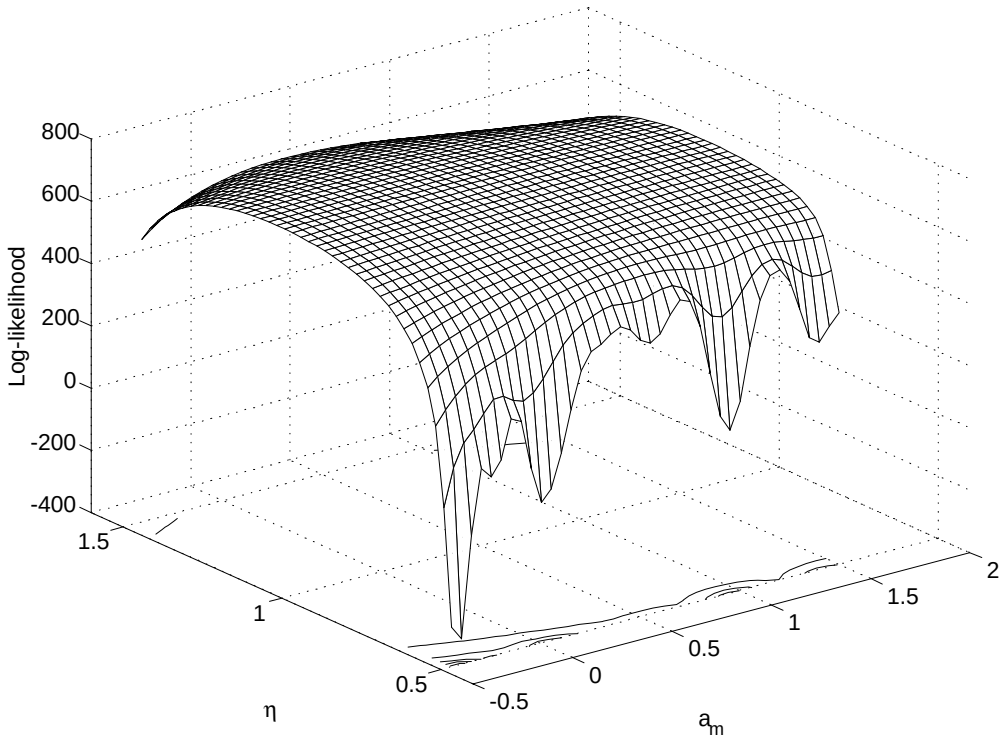


Figure 6.1: Likelihood surface

Note that, at face value, these estimates imply that monetary policy is countercyclical in two specifications and mildly accommodative in the two others. Third, the likelihood function is also relatively flat in the $\sigma_\zeta^2, \sigma_M^2$ space and achieves the maximum around the boundary of the parameter space. Note that with all bivariate data sets, and somewhat counterintuitively, the variance of monetary shocks is estimated to be larger than the variance of the technology shocks. Fourth, the size of the estimated variance of measurement errors is several orders of magnitude larger than the estimated variance of structural shocks, suggesting that misspecification is likely to be present.

Forecasts produced by the model are poor. In fact, the one-step ahead MSEs for hours, productivity growth, output growth and inflation are 30, 12, 7, 15 times larger than the

ones produced by a VAR(1). A test for forecasting accuracy confirms that the forecasts of the model are different from those produced by a VAR(1). The picture improves when a penalty for the larger number of parameters is used. In this case, the value of the Schwartz criterion for the model is "only" twice as large as the one of the VAR(1).

Impulse responses to unitary positive technology and money shocks are in figure 6.2. We report responses obtained with the parameters estimated using productivity and hours data (DATA 1) and output and hours data (DATA 2). Several features are worth discussing. First, estimates of a_M do not affect the responses to monetary shocks. Second, qualitatively speaking, and excluding the responses of output to technology disturbances, the dynamics induced by the shocks are similar across parametrizations. Third, the shape of the responses to technology and monetary shocks looks very similar (up to a sign change) when productivity and hours data are used. Hence, it would be hard to distinguish the two type of shocks by looking at the comovements of these two variables only. Fourth, as expected, the response of productivity to technology shocks is permanent (there is an initial overshooting) and the response of hours is temporarily negative.

Table 6.2 reports cross covariances in the model and in the data. A few features of the table stand out. First, the model estimated with measurement errors fails to capture, both quantitatively and qualitatively, the cross covariance of the data: the magnitude of the estimated covariances is 10 times smaller than the one in the data and the signs of the contemporaneous covariance of $(n_t, \Delta np_t)$, $(\Delta gdp_t, \Delta p_t)$ and $(\Delta np_t, \Delta np_{t-1})$ are wrong.

Second, cross covariances obtained when the model is estimated using productivity and hours data are still somewhat poor. For example, the estimated covariances of $(\Delta gdp_t, \Delta gdp_{t-1})$ and $(\Delta np_t, \Delta np_{t-1})$ are ten times larger than in the data and a distance test rejects the hypothesis that the two set of cross covariances are indistinguishable. Despite these failures, the model estimated using hours and productivity data, captures two important qualitative features of the data: the negative contemporaneous covariance between hours and productivity and the negative lagged covariance of productivity.

Finally, note that neither of the two specifications can reproduce the negative covariance between output growth and inflation found in the data.

Moments/Data	$(\Delta np_t, n_t)$	$(\Delta np_t, n_t, \Delta p_t, \Delta gdp_t)$	Actual data
$\text{cov}(\Delta gdp_t, n_t)$	6.96e-04	4.00e-06	1.07e-05
$\text{cov}(\Delta gdp_t, \Delta np_t)$	5.86e-05	1.56e-06	1.36e-05
$\text{cov}(\Delta np_t, n_t)$	-4.77e-05	1.80e-06	-4.95e-05
$\text{cov}(\Delta gdp_t, \Delta p_t)$	6.48e-04	2.67e-06	-2.48e-05
$\text{cov}(\Delta gdp_t, \Delta gdp_{t-1})$	6.91e-04	3.80e-06	3.443-05
$\text{cov}(\Delta np_t, \Delta np_{t-1})$	-1.51e-04	1.07e-06	-2.41e-05

Table 6.1: Cross covariances

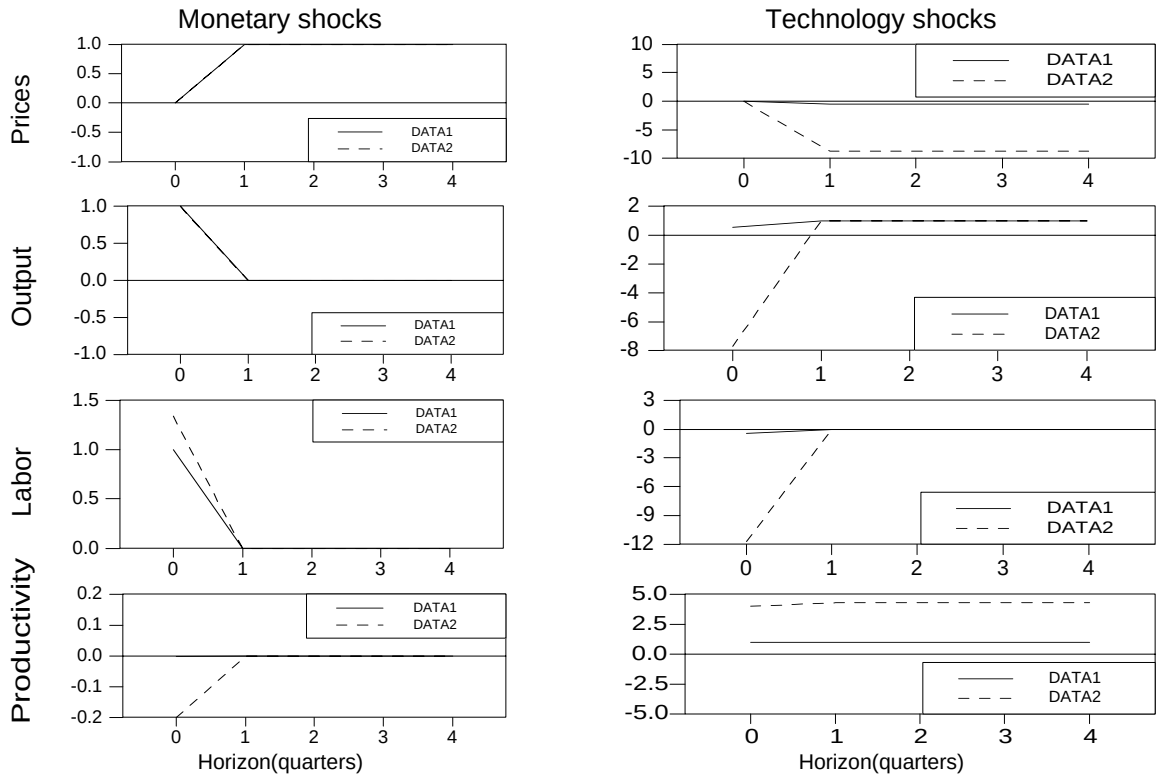


Figure 6.2: Impulse responses

Chapter 7: Calibration

Calibration is an econometric technique which is somewhat different from those we have discussed in the previous two chapters. Given the many alternative definitions existing in the literature, we start by precisely defining what we mean by calibration in this book. As we will see, the approach involves a series of steps which are intended to provide quantitative answers to a particular economic question. From this perspective, a theoretical model is a tool to undertake "computational experiments" rather than a setup to estimate parameters and/or test hypotheses. Also here we take the structure of the model seriously. That is, we start from a formal, abstract, tightly parametrized theoretical construction where the general equilibrium interactions are fully specified. However, contrary to what we have done in the previous two chapters, we will not assume that the model is the data generating process (DGP) for the observables. In fact, one of the basic assumptions of the approach is that theoretical structures used in economic analyses are false in at least two senses: they do not capture all the relevant features of the data; and their probabilistic structure is likely to be misspecified. The reluctance of an investigator to characterize the properties of the discrepancy between the model and the data is also one of the distinctive features of the approach, and this sets it aside from other methodologies, which e.g. assume that the error is a white noise or pay particular attention to its statistical features.

The rest of the chapter describes in details the steps of the procedure, the relationship with more standard estimation/evaluation techniques, with computable general equilibrium exercises and with the type of quantitative exercises conducted in other experimental sciences. Since there is no unifying framework to undertake computational experiments, we organize existing methods according to the way they treat the uncertainty present in various parts of the model. As we will see, approaches can be grouped according to their treatment of sampling, model and other types of uncertainties. Also, we will stress the tight relationship between the methodology used to select the parameters and the one employed to evaluate the quality of the model's approximation to the data. Another useful way to characterize methods is to rank them according to the assumed degree of "falseness" of the model. Different opinions about this features translate in different loss functions and in different criteria to evaluate the magnitude of the discrepancy between the model and the data.

7.1 A Definition

Since the literature has employed the term calibration to indicate different applied procedures (see e.g. Pagan (1994)), confusion may arise when it comes to compare outcomes across methods or studies which only apparently use a similar methodology. For example, it has been suggested that one wants to calibrate a model (in the sense of selecting reasonable parameters values) because there is no data to estimate its parameters. This may be the case when one is interested in quantifying, e.g., the effect of a new tax or of trade liberalization policies in newly created countries. In other cases, such a procedure is employed because the sample is too short to obtain reasonable estimates of a large scale and possible intricate model, or because the data is uninformative about the parameters of interest. In both situations, evaluation of the experimental evidence is problematic. Since all uncertainty is eschewed, back-of-the-envelope calculations are employed to perform some sort of sensitivity analysis on the outcomes of the experiments (see e.g. Pesaran and Smith (1992)). Alternatively, one may prefer to calibrate a model (as opposed to estimated it) if the expected misspecification is so large that statistical estimation of its parameters will produce inconsistent and/or unreasonable estimates and formal statistical testing will lead to outright rejection. Finally, some users interpret calibration as an econometric technique where the parameters are estimated using "economic", as opposed to "statistical", criteria (see e.g. Canova (1994)).

In this chapter, the term calibration is used to indicate a particular collection of procedures designed to provide an answer to economic questions using "false" models. The term "false" is used here in a broad sense: a model is "false" if it approximate the data generating process of (a subset of) the observable data. The essence of the methodology, as stated, e.g. in Kydland and Prescott (1991) and (1996)), can be summarized as follows:

Algorithm 7.1 1) *Choose an economic question to be addressed.*

2) *Select a model design which bears some relevance to the question asked.*

3) *Choose functional forms for the primitives of the model and find a solution for the endogenous variables in terms of the exogenous ones and of the parameters.*

4) *Select parameters and convenient specifications for the exogenous processes and simulate paths for the endogenous variables.*

5) *Evaluate the quality of the model by comparing its outcomes to a set of "stylized facts" of the actual data.*

6) *Propose an answer to the question, characterize the uncertainty surrounding the answer and do policy analyses if required.*

"Stylized facts" is a vague term. Originally, the literature meant a collection of sample statistics which (i) do not involve estimation of parameters and (ii) are easy to compute.

These were typically unconditional moments and, occasionally, conditional moments, histograms or interesting deterministic (nonlinear) functions of the data. More recently, the coefficients of a vector autoregressive model (VAR), the likelihood function or structural impulse responses are taken to be the relevant stylized facts. In step 5) the comparison becomes statistically meaningful only after a measure of distance is selected. This is probably the most important step in the approach: answers obtained from a model that is incapable of explaining observed outcomes are likely to be treated with greater care than those produced by a model which has an excellent record in matching variations in observed series. However, it is also the most controversial part and it is on this issue that most of the methodological debate takes place.

7.2 The Uncontroversial parts

The first two steps of the procedure - choose a question of interest and a model to address it - require little discussion. In general, the questions posed display four types of structures:

- How much of fact X can be explained with impulses of type Y?
- Is it possible to generate features F using theory A?
- Can we reduce the discrepancy D of the theory from the data using feature F?
- How much endogenous variables change if the process for the exogenous variables is altered?

Two questions which have received considerable attention in the literature concern the contribution of technology and/or monetary disturbances to output variability (see e.g. the literature pioneered by Kydland and Prescott (1982) or Chari, Kehoe and McGrattan (2000)) and the ability of a model to quantitatively replicate the excess return of equities over bonds - the so-called equity premium puzzle (see e.g. the literature following Merha and Prescott (1985)). Recently, the literature has investigated the type of frictions needed to reduce discrepancies of certain theories to the data (see e.g. Boldrin, Christiano and Fisher (2001) or Neiss and Pappa (2002)) and examined whether certain policy choices can explain the behavior of real variables in particular historical episodes (e.g. Ohanian (1997), Christiano, Gust and Roldos (2001) or Beaudry and Portier (2002)).

As it is obvious from this extremely incomplete list, the questions posed are clearly specified and the emphasis is on the quantitative implications of the exercise. Occasionally, qualitative implications are also analyzed (e.g. J-curves in the trade balance, see Backus, Kehoe and Kydland (1994)) or humps in responses of certain variables to shocks), but, in general, numerical quantification is the final goal of the exercise.

For the second step - the choice of an economic model - there are essentially no rules: the only requirement being that it has to have some relationship with the question asked. Typically, dynamic general equilibrium models are selected. Both competitive and non-competitive structures have been used (for the latters see Rotemberg and Woodford (1997),

Danthine and Donaldson (1992), or Merz (1995)), and models with fundamental or nonfundamental sources of disturbances have also been studied (see e.g. Farmer (1997)).

It is important to stress that the model one selects is question determined and rarely the modeler attempts to capture all the important features of the data. In other words, one does not expect to have a realistic model to answer important questions. However, to give credibility to her answer, a researcher needs a theory that has been tested through use and found to provide reliable answers to a class of questions. That is to say, a model which has matched reasonably well what happened in previous tax reforms could be a reliable instrument to ask what would happen in a new tax reform. Similarly, a model which captures well features of the real side of the economy could be used to address questions concerning the nominal side.

This observation brings us to an important philosophical aspect of the methodology. In a strict sense, all models are approximations to the DGP and, as such, false and unrealistic. Once this point of view is accepted, it makes no sense to examine the validity of a model using standard statistical tools which assume it to be true, at least under the null. In other words, it is hard to think of a DSGE model as a null hypothesis to be tested - as, for example, it is implicitly assumed with GMM and ML. At best, a DSGE model could be considered an approximation for a subset of the observable data. The problem here has not much to do with limited vs. full information testing - e.g. the Euler equation the model delivers is correct but nothing else is - but with degrees of approximations. What is relevant is the extent a "false" model gives a coherent explanation of interesting aspects of the data. A calibrator is satisfied with her effort if, trough a process of theoretical respecification, the model captures an increasing number of features of the data while maintaining a highly stylized structure. In this sense, the exercises conducted by calibrators belong to the so-called normal science, as described by Kuhn (1970).

Example 7.1 *A closed economy model is probably misspecified to examine, say, the effects of monetary policy disturbances in the EMU: trade with non-EMU countries is about 10-15% of EMU GDP and financial links, especially with the UK, are large. Nevertheless, such a model can be useful to examine the propagation of monetary shocks if it can, for example, replicate the pattern of real responses to monetary shocks and at the same time, say, the responses of real variables to technology disturbances.*

Let y_t be a vector of stochastic processes and $x_t^\dagger = h^\dagger(\epsilon_t, \theta)$ be a model which has something to say about elements of y_t, where ϵ_t are exogenous variables and θ a vector of parameters. Because the model is only an approximation to the DGP of $y_{1t} \subset y_t$, we write

$$y_{1t} = x_t^\dagger + v_t \tag{7.1}$$

where v_t captures the discrepancy between $h^\dagger(\epsilon_t, \theta)$ and the data generating process of y_{1t}. In general, the properties of v_t are unknown. For example, it need not be a mean zero, serially uncorrelated process as it would be the case if $x_t^\dagger = E_t(y_{1t})$ where E_t is the conditional expectations. Evaluating the magnitude of v_t without knowing its properties is virtually impossible.

To be able to provide quantitative answers is necessary to find an explicit solution for the endogenous variables in terms of exogenous and predetermined variables and of the parameters. In this sense, calibration is similar to ML and distinct from GMM, where inference can be conducted without explicit model solutions. We have seen in chapter 2 that analytical solutions can not be obtained, except in very special circumstances. Both local and global approximation procedures generate a $x_t = h(\epsilon_t, \alpha)$ where α is a function of θ and such that $||h - h^\dagger||$ is minimal in some metric. Which of the procedures outlined in chapter 2 one selects depends on the question asked. For example, if the dynamics of the model around the steady state are the focus of the investigation, local approximations are sufficient. On the other hand, in comparing regimes which require drastic changes in the parameters of the control variables, global approximation methods must be preferred.

7.3 Choosing parameters and stochastic processes

With an approximate solution, paths for the endogenous variables can be obtained once the vector θ and the properties of ϵ_t are specified. The selection of the properties of ϵ_t is relatively uncontroversial. One either chooses specifications which are tractable, e.g. an AR process with arbitrary persistence and innovations which are transformations of a $\mathsf{N}(0, 1)$ process, or that give some realistic connotation to the model, e.g. select the Solow residuals of the actual economy, the actual path of government expenditure or of the money supply, with the second alternative being preferred if policy analyses are undertaken. There is more controversy, on the other hand, when it comes to select the parameters of the model. Typically, they are chosen so that the model reproduces certain observations. The next example clarifies how this is done and implicitly explains why calibration is at times referred to as "computational experiment".

Example 7.2 *Consider the problem of measuring the temperature of the water in various conditions. To conduct the experiment an investigator will have to set the measurement instrument (in this case, a thermometer) to insure that the outcome is accurate. One way of doing this is to graduate the thermometer to some observations. For example, if the experiment consists in measuring at what temperature the water boils on the top of a mountain, a researcher could, at sea level, set the tick corresponding to freezing water to zero and the one corresponding to boiling water to 100, interpolate intermediate values with a linear scale and use the graduated thermometer to undertake the measurement. Alternatively, if the experiment consists in checking the amount of heat released by a boiler over a period of time, she can use the calibrated thermometer to check the water temperature after the machine has been operating for, say, 5, 10 and 20 minutes.*

In a way, the process of selecting the parameters an economic model is similar. A model is an instrument which needs to be graduated before the measurement of interest is performed. There are at least two ways of doing this graduation: the one suggested in the computable general equilibrium (CGE) tradition summarized, e.g. in Showen and Walley

(1984)-(1992) and the one used in modern DSGE models, see, e.g. Kydland and Prescott (1982). While similar in spirit, the two methodologies have important differences.

In CGE models, a researcher typically solves a large, nonlinear intersectorial static model, linearizing the system of equations around an hypothetical equilibrium where prices and quantities clear the markets. It is not necessary that this equilibrium exists. However, because the coefficients of the linear equations are functions of equilibrium values, it is necessary to measure it. CGE users need to find a "benchmark data set" and make sure that the linearized model replicates this data. Finding such a data set is complicated and requires ingenuity. Often, the selection process leaves some of the parameters undetermined. In this case a researcher assigns them arbitrary values or fixes them using existing estimates (e.g. estimates obtained in countries at similar stages of development) and then performs rough sensitivity analysis to determine how the outcomes vary when these parameters are changed. Although the procedure to select the free parameters and the way sensitivity analysis is undertaken are arbitrary, the procedure is coherent with the philosophy of CGE models: a researcher is interested in examining deviations from an hypothetical equilibrium not from an economy in real time (see e.g. Kim and Pagan (1994) for a discussion).

In DSGE models, the equilibrium the model needs to reproduce is typically the steady state or, in case of models with frictions, the Pareto optimal equilibrium. In the former case, parameters are chosen so that the steady state for the endogenous variables replicates time series averages of the actual economy. In the latter case, parameters are selected so that the model without frictions matches certain features of the actual data. Also in this case, the chosen conditions do not pin down all the parameters and different researchers have used different techniques to choose the remaining ones. For example, one can select these parameters a-priori; pin them down using available estimates; informally estimate them with a method of moment or formally estimate them using GMM (see e.g. Christiano and Eichenbaum (1992)), SMM (see e.g. Canova and Marrinan (1993)) or ML (see e.g. McGrattan, Rogerson and Wright (1993)) procedures. However, choosing the parameters with one of these latters three approaches is inconsistent with the philosophy of the methodology, since the dimensions used to estimate the free parameters can no longer be considered approximations to the DGP.

Formally, let $\theta = (\theta_1, \theta_2, \theta_3)$, let θ_1 be the parameters which appear in the equilibrium conditions and θ_2, θ_3 two sets of free parameters. In CGE models θ_3 are absent while $\theta_1 = h_1(y_0, \epsilon_0, \theta_2) \equiv h_1(\theta_2)$ where (y_0, ϵ_0) are the hypothetical data. Then $y_{1t} = h(\epsilon_t, \theta_1, \theta_2) \equiv \tilde{h}(\epsilon_t, \theta_2)$. Hence, if ϵ_t is deterministic, the range of y_{it} to variations in θ_2 can be calculated using the numerical derivatives of $\tilde{h}$ i.e. obtain $\frac{\tilde{h}(\theta_2+\iota)-\tilde{h}(\theta_2-\iota)}{2\iota}$, $\iota > 0$ and small. This can be done informally (trying few values), conditionally (perturbing one parameters at the time or using a grid), or formally (linearizing $\tilde{h}$ and using asymptotic theory). Also in DSGE models, given θ_2, $\theta_1 = h_1(y_0, \epsilon_0, \theta_2)$ so that $y_{1t} = \tilde{h}(\epsilon_t, \theta_2, \theta_3)$. However, here θ_3 are selected to minimize some quantity, e.g. $[\mathsf{S}(\frac{1}{T}\sum_t y_t y_{t-\tau}) - \mathsf{S}(\tilde{h}(\epsilon_t, \theta_2, \theta_3)h(\epsilon_{t-\tau}, \theta_2, \theta_3)')]$ for some τ, where S is a selection matrix, either informally or formally.

Example 7.3 *(Selecting the parameters of a RBC model) Suppose that the social planner*

maximizes $E_0 \sum_t \beta^t \frac{(c_t^{\vartheta}(1-N_t)^{1-\vartheta})^{1-\varphi}}{1-\varphi}$ *by choices of* (c_t, K_{t+1}, N_t) *subject to*

$$G_t + c_t + K_{t+1} = \zeta_t K_t^{1-\eta} N_t^{\eta} + (1-\delta)K_t \equiv GDP_t + (1-\delta)K_t \tag{7.2}$$

where $\ln \zeta_t = \bar{\zeta} + \rho_\zeta \ln \zeta_{t-1} + \epsilon_{1t}, \quad \epsilon_{1t} \sim (0, \sigma_\zeta^2),\ \ln G_t = \bar{G} + \rho_g \ln G_{t-1} + \epsilon_{4t}, \quad \epsilon_{4t} \sim (0, \sigma_g^2)$, K_0 *given, where* c_t *is consumption,* N_t *is hours worked,* K_t *is the capital stock. We assume that* G_t *is financed with lump sum taxes or bond creation. Letting* λ_t *be the Lagrangian on (7.2), the log linearized optimality conditions are*

$$\begin{aligned}
\hat{\lambda}_t - (\vartheta(1-\varphi)-1)\hat{c}_t + (1-\vartheta)(1-\varphi)\frac{N^{ss}}{1-N^{ss}}\hat{N}_t &= 0 && (7.3)\\
\hat{\lambda}_{t+1} + \frac{(1-\eta)(GDP/K)^{ss}}{(1-\eta)(GDP/K)^{ss} + (1-\delta))}(\hat{GDP}_{t+1} - \hat{K}_{t+1}) &= \hat{\lambda}_t && (7.4)\\
\frac{1}{1-N^{ss}}\hat{N}_t + \hat{c}_t - \hat{GDP}_t &= 0 && (7.5)\\
\hat{w}_t - \hat{GDP}_t + \hat{n}_t &= 0 && (7.6)\\
\hat{r}_t - \hat{GDP}_t + \hat{k}_t &= 0 && (7.7)\\
\hat{GDP}_t - \hat{\zeta}_t - (1-\eta)\hat{K}_t - \eta\hat{N}_t &= 0 && (7.8)\\
(\frac{G}{GDP})^{ss}\hat{G}_t + (\frac{c}{GDP})^{ss}\hat{c}_t + (\frac{K}{GDP})^{ss}(\hat{K}_{t+1} - (1-\delta)\hat{K}_t) - \hat{GDP}_t &= 0 && (7.9)
\end{aligned}$$

where the first three equations come from consumers decisions; the next two from firms decisions; the last two represent the production function and the resource constraint and the superscript SS *indicates steady state values. Here* K_t *is the state,* (ζ_t, G_t) *the shocks, and there are 6 endogenous variables* $(\hat{\lambda}_t, \hat{c}_t, \hat{N}_t, \hat{GDP}_t, \hat{w}_t, \hat{r}_t)$. *Since there are seven equations and seven unknowns a solution exists. The model has four types of parameters:*

(i.) Technological parameters (η, δ).

(ii.) Preference parameters $(\beta, \vartheta, \varphi)$.

(iii.) Steady state parameters $(N^{ss}, (\frac{c}{GDP})^{ss}, (\frac{K}{GDP})^{ss}, (\frac{G}{GDP})^{ss})$.

(iv.) Auxiliary (nuisance) parameters $(\bar{\zeta}, \bar{g}, \rho_g, \rho_\zeta, \sigma_\zeta^2, \sigma_g^2)$.

Equations (7.3)-(7.5) and (7.9) imply in the steady state:

$$\begin{aligned}
\frac{1-\vartheta}{\vartheta}(\frac{c}{gdp})^{ss} &= \eta\frac{1-N^s}{N^{ss}} && (7.10)\\
\beta[(1-\eta)(\frac{GDP}{K})^{ss} + (1-\delta)] &= 1 && (7.11)\\
(\frac{G}{GDP})^{ss} + (\frac{c}{GDP})^{ss} + \delta(\frac{K}{GDP})^{ss} &= 1 && (7.12)
\end{aligned}$$

(7.10)-(7.12) determine e.g. $(N^{ss}, (\frac{c}{GDP})^{ss}, (\frac{K}{GDP})^{ss})$ *once* $(\frac{G}{GDP})^{ss}, \beta, \vartheta, \eta$ *and* δ *are selected (the formers play the role of* θ_1 *and the latters the role of* θ_2 *in this example). The remaining free parameters can be selected as follows. The production function can be used to provide an estimate of* ζ_t *from which estimates of* $\bar{\zeta}, \rho_\zeta, \sigma^2_\zeta$ *can be backed out. Data for government expenditure can be used to back out the parameters of the* G_t *process. For the last preference parameter, one can appeal to estimates obtained in other studies noting e.g. that coefficient of relative risk aversion is* $1-\vartheta(1-\varphi)$*; could fix it at some arbitrary value; use the Euler equation and the intratemporal condition to get e.g. a GMM estimate; or select it using simulation estimators (for example, choose* φ *so that consumption variability in actual data is the same as in simulated data).*

Note that the log linearized conditions in (7.3)-(7.9) have the form of a vector autoregression of order 1 or of a state space system. Letting $y_t = (\hat{\lambda}_t,\ \hat{K}_t,\ \hat{c}_t,\ \hat{N}_t,\ \hat{GDP}_t,\ \hat{w}_t,\ \hat{r}_t)$*, the VAR representation is* $\mathcal{A}_0 y_{t+1} = \mathcal{A}_1 y_t + \mathcal{A}_2 E_t$ *where* $E_t = [\hat{\zeta}_t, \hat{G}_t]'$ *and*

$$\mathcal{A}_0 = \begin{bmatrix} 1 & -\frac{(1-\eta)\frac{GDP}{K}^{ss}}{(1-\eta)(\frac{GDP}{K})^{ss}+(1-\delta)} & 0 & 0 & \frac{(1-\eta)(\frac{GDP}{K})^{ss}}{(1-\eta)\frac{GDP}{k}^{ss}+(1-\delta)} & 0 & 0 \\ 0 & 1/(\frac{GDP}{K})^{ss} & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}, \mathcal{A}_2 = \begin{bmatrix} 0 & 0 \\ 0 & -(\frac{G}{GDP})^{ss} \\ 0 & 0 \\ 0 & 0 \\ -1 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}$$

$$\mathcal{A}_1 = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & (1-\delta)/(\frac{GDP}{K})^{ss} & -(\frac{c}{GDP})^{ss} & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & \frac{1}{1-N^{ss}} & -1 & 0 & 0 \\ -1 & 0 & \vartheta(1-\varphi)-1 & -(1-\vartheta)(1-\varphi)\frac{N^{ss}}{1-N^{ss}} & 0 & 0 & 0 \\ 0 & 1-\eta & 0 & \eta & -1 & 0 & 0 \\ 0 & 0 & 0 & 1 & -1 & 1 & 0 \\ 0 & 1 & 0 & 0 & -1 & 0 & 1 \end{bmatrix}$$

A state space representation is obtained setting $y_{2t} = (\hat{\lambda}_t, \hat{k}_t, \zeta_t, G_t)$; $y_{1t} = (\hat{c}_t, \hat{N}_t, \widehat{gdp}_t, \hat{w}_t, \hat{r}_t)$*,* $y_{3t} = (\epsilon_{1t}, \epsilon_{4t})$*:*

$$\begin{aligned} \mathcal{A}_0^2 y_{2t+1} &= \mathcal{A}_1^2\ y_{2t} + \mathcal{A}_2^2 y_{3t} \\ \mathcal{A}_0^1 y_{1t+1} &= \mathcal{A}_1^1\ y_{2t} + \mathcal{A}_2^1 y_{3t} \end{aligned} \tag{7.13}$$

where $\mathcal{A}_0^i, \mathcal{A}_1^i, \mathcal{A}_2^i, i = 1, 2$ *are appropriate partitions of* $\mathcal{A}_0, \mathcal{A}_1, \mathcal{A}_2$*. Since* y_{1t} *does not enter in the equation for* y_{2t}*, it does not Granger cause* y_{2t} *(see chapter 4).*

Exercise 7.1 *Consider a sticky price model without capital and instantaneous utility* $U(c, N, M) = \ln c_t + \frac{1}{1-\varphi_m}(\frac{M_{t+1}}{p_t})^{1-\varphi_m}$*. Assume Calvo pricing; let* $1-\zeta_p$ *be the fraction of agents allowed to change prices and* β *the discount factor. Derive an Euler equation, a money demand function and a Phillips curve. Log-linearize the conditions and describe how to select preferences and production parameters, relevant steady states and auxiliary parameters.*

Exercise 7.2 *Consider adding capacity utilization to the model of example 7.3. That is, assume that the production function depends on both capital K_t and its utilization ku_t and it is of the form $GDP_t = \zeta_t(K_t ku_t)^{1-\eta} N_t^{\eta}$. Assume also that depreciation is related to utilization via the equation $\delta(ku_t) = \delta_0 + \delta_1 ku_t^{\delta_2}$ where δ_0, δ_1 and δ_2 are parameters. Describe how to select $(\delta_0, \delta_1, \delta_2)$.*

Exercise 7.3 *Consider the two country model described in example 2.3 of chapter 2. Log-linearize the first order conditions, the budget constraint and the definitions of the terms of trade (TOT_t) and net export (nx_t). Describe how to choose the free parameters.*

Because not all parameters can be pinned down by the reference equilibrium, there is a degree of arbitrariness inherent in the procedure. Furthermore, all approaches designed to choose parameters not appearing in the steady state have advantages and disadvantages. For example, employing information present in existing studies has the advantage of allowing a researcher to pin down parameters which can not be identifiable from the available data. However, a selectivity bias is typically present (see Canova (1995)): there is a variety of estimates available and different researchers may refer to different studies even when examining the same question. Furthermore, such an approach artificially reduces the uncertainty surrounding the predictions of the model and this may generate an unwarranted confidence in the outcomes of the experiment. Finally, inference may be spurious and/or distorted. In fact, estimates of θ_3 may be biased and inconsistent unless the selected θ_2 are the true parameters of the DGP or consistent estimates of them.

Example 7.4 *To illustrate this latter problem consider a framework typically studied in undergraduate econometric textbooks where there is a linear relationship between x and y and the disturbance is serially correlated. Letting α be the parameters of the linear relationship and ρ the AR(1) coefficient of the errors, a GLS estimator for (α, ρ) is $y_t - \rho y_{t-1} = \alpha(x_t - \rho x_{t-1}) + e_t$ where $e_t \sim$ iid $(0, \sigma_e^2)$. An estimator for α, conditional on ρ, is $(\hat{\alpha}|\rho) = ((x_t - \rho x_{t-1})'(x_t - \rho x_{t-1}))^{-1}(x_t - \rho x_{t-1})'(y_t - \rho y_{t-1}))$ If ρ is consistently estimated, $(\hat{\alpha}|\hat{\rho}) \xrightarrow{P} (\hat{\alpha}|\rho)$ as $T \to \infty$. However, if it is not the case, the asymptotic distribution of $(\hat{\alpha}|\hat{\rho})$ will be centered around a wrong value. In table 7.1 we verify that biases do occur: we report the mean and the interquartile range of the Monte Carlo distribution of $(\hat{\alpha}|\rho)$ obtained conditioning on $\rho = 0.0, 0.4, 0.9$, when $T = 1000$, 1000 replications are used to construct distributions, and the true values are $\alpha = 0.5$ and $\rho = 0.9$.*

	25th percentile	mean	75th percentile
$\rho = 0$	0.396	0.478	0.599
$\rho = 0.4$	0.443	0.492	0.553
$\rho = 0.9$	0.479	0.501	0.531

Table 7.1: Monte Carlo distribution of $\alpha|\rho$

While example 7.4 is highly stylized, Gregory and Smith (1989) have shown that this problem could be important, for example, in determining the range of the risk aversion parameter which is consistent with the magnitude of US equity premium.

Exercise 7.4 *(Risk free rate puzzle) Take the economy discussed in exercise 2.6 of chapter 2. Assume that* $u(c_t) = \frac{c_t^{1-\varphi}}{1-\varphi}$ *and that output evolves according to* $gdp_{t+1} = gy_{t+1}gdp_t$. *Assume that* gy_{t+1} *can take* n *possible values* $(gy_1, \ldots, gy_n)$ *and let* $p_{ij} \equiv P(gy_{t+1} = gy_{i'}|gy_t = gy_i) = \mu_{i'} + \rho_y(\mathcal{I}_{ii'} - \mu_{i'})$ *where* $\mu_{i'}$ *are unconditional probabilities,* $\mathcal{I}_{ii'} = 1$ *if* $i = i'$ *and zero otherwise, and* $\rho_y \in (-(n-1)^{-1}, 1)$. *If the current state is* (gdp_t, gy_i), *the price of an asset paying one unit of output next period satisfies* $p_t^s U_c(gdp_t, gy_i) = \sum_{i'} p_{ii'}\beta[U_c(gdp_{t+1}, gy_{i'})]$. *It is easy to verify that unconditionally* $var(p_t^s) = \beta^2\rho_y^2 \sum_{i=1}^n p_i(\sum_{i'=1}^n p_{i'}(gy_i^{-\varphi} - gy_{i'}^{-\varphi}))^2$.
i) Set $n = 2, gy_1 = 0.9873, gy_2 = 1.0177, \mu_1 = 0.2, \mu_2 = 0.8, \beta = 0.99, \rho_y = 0.8, \varphi = 2$. *Simulate asset price data from the model and treat these as the actual data.*
ii) Set (n, gy_i, p_i, β) *as in i) but now set* $\rho_y = 0.6$. *Choose* φ *so that the simulated variance of asset prices matches the variance of asset prices produced in i) (you can select the loss function you want, e.g.,* $min|var^A(p_t^s) - var^S(p_t^s)|$ *where* $var^S(p_t^s)$ $(var^A(p_t^s))$ *is the variance of simulated (actual) data). Repeat the exercise for* $\rho_y = 0.9$.
iii) Repeat i) 100 times drawing φ *from* $\mathsf{U}(1, 10)$ *(treat this as 100 realizations of the actual data). Repeat ii) 100 times for* $\rho_y = 0.6, 0.9$ *and show the distributions of* $\hat{\varphi}$ *that best matches* $var^A(p_t^s)$.
iv) Consider now the mean of asset prices. Repeat iii) fixing $\varphi = 2$ *and choosing* ρ_y *to minimize* $|E_t((p_t^s)^A - (p_t^s)^S)|$. *Is there any pattern worth mentioning?*

As pointed out by Kydland and Prescott (1991), choosing parameters using information obtained from other studies imposes coherence among various branches of the profession. For example, one uses growth models to examine business cycles fluctuations and checks its implications using parameters obtained from e.g. micro studies of labor markets. However, for many parameters, available estimates are surprisingly sparse (see e.g. Showen and Walley (1992, p.105)) and often they are obtained with estimation procedures which, although valid in the environment where they were produced, make no sense in DSGE frameworks (see Hansen and Heckman (1996)). Hence, any choice is arbitrary and sensitivity analysis is needed to evaluate the robustness of the measurement made to changes in these parameters.

Canova (1994)-(1995) suggested an approach which responds to these criticisms. Instead, of fixing θ_3 to one particular value, he restricts its range to an interval using theoretical considerations and uses all the information to construct an empirical distribution for θ_{3i}, $i = 1, 2, \ldots$ over this interval (this is treated as the likelihood of a parameter, given existing estimates) and draws for θ_3 are made from there joint "empirical" distribution. Moreover, since the distinction between θ_2 and θ_3 is artificial, intervals for both sets of parameters are typically used (see e.g. Canova and Marrinan (1996) or Maffezzoli (2000)). An example may clarify the approach.

Example 7.5 *In the exercise 7.4 one of the free parameters is the coefficient of constant relative risk aversion* φ. *Typically one sets* φ *to 1 or 2 (resulting in a mild curvature of*

the utility function) and, occasionally, tries a few larger values to construct upper bound measures. As an alternative, one could limit the range of values using economic arguments to, say, [0,20] and construct a smoothed histogram over this interval using existing estimates. Since most estimates are in the range [1,2] and since in some asset pricing models researchers have tried values up to 10, the empirical distribution for φ could be approximated with a $\chi^2(4)$, which has the mode at 2 and about 5% probability in the region above 10. When no empirical information exists, one chooses a uniform distribution or a distribution capturing the subjective beliefs a researcher has about the likelihood of the parameter (see chapter 9).

Exercise 7.5 *Suppose the representative agent maximizes* $E_o \sum_t \beta^t (\frac{c_t^{1-\varphi}}{1-\varphi} + \frac{\vartheta_M (M_{t+1}/p_t)^{1-\varphi_m}}{1-\varphi_m})$ *subject to* $c_t + K_t + \frac{M_{t+1}}{p_t} = \zeta_t K_{t-1}^{1-\eta} N_t^{1-\eta} + (1-\delta)K_{t-1} + \frac{M_t}{p_t}$ *where* $\ln M_t = \bar{M} + \rho_M \ln M_{t-1} + \epsilon_{3t}$, $\ln \zeta_t = \bar{\zeta} + \rho_\zeta \ln \zeta_{t-1} + \epsilon_{1t}$ *and* $\epsilon_{1t}, \epsilon_{3t}$ *are iid with standard error equal to* σ_ζ, σ_M. *Describe what distributions you would choose for the parameters* $(\eta, \varphi, \delta, \vartheta_M, \varphi_M, \rho_\zeta, \rho_g, \sigma_\zeta, \sigma_M)$.

Standard statistical estimation of the free parameters has three main advantages: it avoids arbitrary choices; it provides a coherent framework for choosing *all* the parameters; and produces measures of uncertainty which can be used to evaluate the quality of the approximation of the model to the data. The disadvantages are of various kinds. First, formal or informal procedures require the selection of the moments/statistics to be matched, and that may lead to inconsistencies across studies. The method employed to select θ_3 in example 7.3 can indeed be thought as a method of moment estimation where parameters are chosen to match the first moments of the data (i.e. the long run averages). Christiano and Eichenbaum (1992); Feve and Langot (1994) and others, use first and second moments of actual and simulated data to obtain GMM estimates of the parameters. Smith (1993) uses the scores of the likelihood. While the selection of moments depends on the question asked, efficient estimation requires all moments containing information on the parameters to be used (see chapter 5). Second, GMM estimates are biased in small (or nonstationary) samples (see chapter 5). Therefore, simulations conducted with these estimates may lead to spurious inference. Third, informal SMM approaches may produce estimates of parameters even though they are not identifiable. One example of this phenomena was provided in point (iv) of exercise 7.4. Finally, one should note that the type of uncertainty present in the outcomes of the model when parameters are estimated is different from the uncertainty existing when a calibrator is ignorant about the magnitude of a parameter. In fact, once the data and moments are selected, sample uncertainty is typically small. Since the measurement depends only on y_{1t} and on the estimator chosen, uncertainty in measurement is also small. However, the uncertainty present in e.g. choosing a risk aversion parameter, is typically large.

As mentioned in chapter 5, ML estimation can be thought as a GMM procedure where the moments are the scores of the likelihood function. Therefore, discrepancies between estimates obtained with these two methods may indicate either that either the orthogonality conditions span a different informational space (GMM may use only parts of the model while ML uses all of it) or that the sample strongly deviates from normality. Asymptotically, when

the moment restrictions and the scores span the same space, the two procedures must give identical results. Hence, all the arguments made for GMM or SMM also apply to ML.

It is useful to compare the parameter selection process used by a calibrator with the one used in a traditional econometric approach. In the latter parameters are chosen to minimize some statistical criteria, e.g., the MSE. Such a loss function however does not have any economic justification, its conventional use reflects mathematical convenience and imposes stringent requirements on the structure of v_t. The loss function used by calibrators, on the other hand, has an economic interpretation: parameters are chosen so that the steady state of the model matches the long run averages of the data. However, since not all parameters are pinned down by these conditions, a calibrator may look like an econometrician who uses different loss functions in different parts of the model. Furthermore, since choosing parameters to match long run observations is equivalent to using GMM on first moments only, a calibrator may also look like as an inefficient GMM econometrician.

Finally, note that when intervals are chosen for the free parameters and empirical distributions are used, the parameter selection procedure shares a tight connection with Bayesian methods, which we will discuss in details in chapters 9 to 11 of this book.

7.4 Model Evaluation

Before the measurement of interest is undertaken it is necessary to assess the quality of the model's approximation to the data. The most active branch of the literature is concerned with the development of methods to evaluate the fit of calibrated models. Classical pieces, such as Kydland and Prescott (1982), are silent on this issue. But this is not completely surprising: since there are no free parameters and no uncertainty is allowed in either the selected parameters or the moments used for comparison, the model deterministically links the endogenous variables to the parameters and exogenous stochastic processes. Hence, unless the sampling variability of the exogenous processes is used, measures of distance between the model and the data can not be defined. The lack of formal model validation does not seem to bother some researchers. Kydland and Prescott (1991), (1996) for example, emphasize that the trust a researcher puts in an answer given by the model does not depend on statistical measure of discrepancy, but on how much she believes in the economic theory used and in the measurement undertaken - in other words, trust could be an act of faith.

Nowadays, most calibrators informally compare the properties of simulated data to a set of stylized facts of the actual data. Such an approach is in fashion also with econometric skeptics: simple sample statistics are believed to be sufficient to do the job since "either you see it with naked eyes or no fancy econometrics will find it". The choice of stylized facts obviously depends on the question asked but one should be aware that there are many ways to summarize the outcome of a calibration exercise and some may be more informative than others for comparison purposes.

In a business cycle context one typically selects a subset of auto and cross-covariances of the data, but there is no reason for focusing on unconditional second moments, except that their measurement does not require the estimation of time series models. One could

also use the distributions of actual and simulated data - which also do not need parametric time series models to be estimated - or their VAR representation and examine some of their statistical features (e.g. the number of unit roots or exclusion restrictions (as in Canova, Finn and Pagan (1994)), the magnitude of VAR coefficients (as in Smith (1993) or Ingram, DeJong and Whiteman (1996)), or the pattern of semi-structural impulse responses (as in Cogley and Nason (1994)). Alternatively, one could reduce the model to one or two equations and compare the time series representations of the variables in the model and in the actual data (as in Canova, Finn and Pagan (1994) or Cogley and Nason (1995)). Finally, business cycle turning points (as in King and Plosser (1994) or Simkins (1994)), variance bounds (as in Hansen and Jagannathan (1991)), durations and asymmetries of the cycle (as Pagan and Harding (2002) or historical episodes (as in Ohanian (1997) or Beaudry and Portier (2001)), could also be used for evaluation purposes.

Example 7.6 *(Magnitude of the VAR coefficients/ exclusion restrictions)*
A log linearized DSGE model has a state space representation where the endogenous variables depend on the first lag of the states and on the shocks. For example, in the RBC model of example 7.3 we have $\mathcal{A}_0^1 y_{1t} = \mathcal{A}_1^1 y_{2t-1} + A_2^1 y_{3t}$ *where the matrices* $\mathcal{A}_i^1, i = 0, 1, 2$ *are functions of the "deep" parameters* θ *of the model. Hence, once these are selected, they are matrices of real numbers. A simple RBC model then poses two types of restrictions on the VAR of say, output, consumption, investment and hours. First, lagged values of these four variables should not help to predict current values once lagged values of the states are included. Second, in a regression of* y_{1t} *on* y_{2t-1}*, the coefficient matrix must be equal to* $(\mathcal{A}_0^1)^{-1}\mathcal{A}_1^1$.

Example 7.7 *(Final form comparison) The RBC model described in example 7.3 can also be reduced to a bivariate VARMA(1,1) for* $(\hat{N}_t, \hat{c}_t)$*. Solving for* $\hat{c}_t$*,* $\hat{N}_t$ *has an* $ARMA(\infty, \infty)$ *representation of the type* $A(\theta)(\ell)\hat{N}_{t+1} = D(\theta)(\ell)e_{t+1}$*, where the reduced form parameters* $A(\ell), D(\ell)$ *are functions of the "deep" parameters* θ *and* $\hat{e}_t = (\hat{\zeta}_t, \hat{g}_t)$*. Given this representation there are at least two ways of comparing the data and the model. First, one can compare the autocorrelation function of hours produced by the model (conditional on* $\theta = \bar{\theta}$*) with the autocorrelation function of hours found in the data. Second, we can estimate an ARMA model and verify whether (i) an* $ARMA(\infty, \infty)$ *fits the data, (ii) the estimated coefficients are exactly equal to those implied by the model.*

Table 7.2 reports few terms of the ACRF of a the version of the model where $u(c_t, N_t) = \ln(c_t) + \vartheta_N(1 - N_t)$*, there is no government and* $\beta = 0.99, \eta = 0.64, \vartheta_N = 2.6, \delta = 0.025, \rho_\zeta = 0.95, \sigma_\zeta^2 = 0.007$*; the same ACRF terms obtained from linearly detrended US data (using Seasonally Adjusted Average Weekly Hours of Private Nonagricultural Establishments for the sample 1964:1-2003:1) and estimates of the best ARMA specification obtained in the data. In parenthesis are standard errors. It is clear that while standard deviations are similar, the model's ACRF function is less persistent than the data's. In fact, the 12th order correlation in the data is still 0.786, while it is roughly zero in the model. Moreover, an ARMA(2,2) only partially fits US data: for example, neither the AR(2) nor the MA coefficients are significant but a Q-test shows the presence of residual autocorrelation (pre-*

sumably, there are higher order dynamics in the data). Finally, the estimated values are significantly different from those implied by the model (the AR(1) and AR(2) coefficients in the model are, respectively, 1.57 and -0.53). While it is obvious that the model fails to captures the dynamics of actual hours, it is hard to see how to reduce the discrepancy. Lack of persistence could be due to many reasons (lack of investment propagation, lack of intertemporal substitutability, etc.) and the reduced form approach used here does not allow us to disentangle them.

	Standard deviation	$\text{Corr}(N_t, N_{t-1})$	$\text{Corr}(N_t, N_{t-2})$	$\text{Corr}(N_t, N_{t-3})$
Actual data	0.517(0.10)	0.958(0.09)	0.923(0.09)	0.896(0.09)
Simulated data	0.473	0.848	0.704	0.570
	Estimated ARMA(2,2) for actual hours			
	AR(1)	AR(2)	MA(1)	MA(2)
Actual data	1.05(0.24)	-0.07 (0.21)	-0.12 (0.21)	-0.05(0.09)

Table 7.2: ACF of hours

Exercise 7.6 *Consider the cash-in-advance model of example 2.4 of chapter 2 where all consumption goods are cash goods, the representative agent maximizes* $E_0 \sum_t \beta^t \frac{c_t^{1-\varphi}}{1-\varphi}$ *subject to* $p_t c_t \leq M_t + T_t$ *and* $c_t + K_{t+1} + \frac{M_{t+1}}{p_t} \leq r_t K_t + (1-\delta)K_t + \frac{M_t + T_t}{p_t}$ *where* $T_t = M_{t+1} - M_t$, $\ln M_{t+1} = \bar{M} + \rho_m \ln M_t + \epsilon_{3t}$, $\epsilon_{3t} \sim (0, \sigma_m^2)$. *Assume the production function* $GDP_t = \zeta_t K_t^{1-\eta}$ *where* $\ln \zeta_t$ *is an AR(1) with persistence* ρ_ζ *and variance* σ_ζ^2.
(i) Derive a trivariate log-linear (final form) representation for (c_t, M_t, p_t).
(ii) Using US data on consumption, M1 and CPI estimate a trivariate VAR compare the magnitude of VAR coefficients and of the auto and cross correlation function of M1 growth and of consumption growth in the model and in the data (Hint: the model is a VAR(∞*)).*

Exercise 7.7 *Consider the RBC model described in example 7.3 but now assume preferences are given by* $u(c_t, c_{t-1}, N_t) = \frac{(c_t^\gamma c_{t-1}^{1-\gamma})^{1-\varphi}}{1-\varphi} + \vartheta_N(1 - N_t)$, *where* ϑ_N *is a constant. Log linearize and appropriately select* $(\beta, \varphi, \gamma, \eta, \delta, \vartheta_N)$, *the parameters governing the stochastic process for* ζ_t, G_t *and steady state ratios and simulate data. Define an upturn as the situation where* $gdp_{t-2} < gdp_{t-1} < gdp_t > gdp_{t-1} > gdp_{t-2}$ *and a downturn as a situation where* $gdp_{t-2} > gdp_{t-1} > gdp_t < gdp_{t-1} < gdp_{t-2}$. *Examine whether the model matches the turning points of US output using one realization of technology and of government disturbances.*

Exercise 7.8 *Using the sticky price model of exercise 7.1 and the selected parameters, examine whether the model reproduces the persistence of US inflation by computing* $\mathcal{S}(\omega = 0) = \sum_{\tau=-\infty}^{\infty} ACF_\pi(\tau)$ *where* $ACF_\pi(\tau)$ *is the autocovariance of inflation at lag* τ.

At times, it may be more relevant to know how good a model is not in absolute terms, but relative to other competitors. Such a "horse race " is important, for example, when

two models "poorly" approximate the data or when one model is a restricted version of the other. Canova, Finn and Pagan (1994), for example, take the capacity utilization model of Burnside, et al. (1993), reduce it to two equations involving output and investment and compare its performance to a simple investment accelerator model. More recently, using the techniques described in chapters 9 to 11, Schorfheide (2000), DeJong, Ingram and Whiteman (2000) or Smets and Wouters (2003) have undertaken similar comparisons.

Exercise 7.9 *Consider two variants of the RBC model of example 7.3. In variant i) assume that there are production externalities, i.e.* $y_{it} = \zeta_t \bar{K}_t^{\aleph} K_{it}^{1-\eta} N_{it}^{\eta}$ *where* $\bar{K}_t = \int_0^1 K_{it} di$ *is the aggregate capital stock. In variant ii) assume one period labor contracts, so that* $w_t = E_{t-1} \frac{y_t}{N_t}$*. Suggest ways to compare the relative performance of the two models to the data.*

Comparisons based on stylized facts are important for two reasons. First, they shift the emphasis away from statistical quantities (such as the properties of the residuals) towards more interesting economic objects (such as functions of conditional and unconditional moments). Second, they allow to construct a larger set of diagnostics and therefore a better comparison of the properties of different models. Within this generic comparison methodology, several variants, closely linked to the procedure used to select the parameters, are available (see e.g. Kim and Pagan (1993)).

Formally, let S_y be a set of interesting economic statistics of the actual data and let $\mathsf{S}_x(\epsilon_t, \theta)$ be the corresponding statistics of simulated data, given a vector of parameters θ and a vector of driving forces ϵ_t. Model evaluation consists in selecting a loss function L measuring the distance between S_y and S_x and assessing its magnitude. At the cost of oversimplifying, but hopefully gaining a clearer understanding of the differences, we divide existing procedures into four groups:

- Approaches based on R^2-type measure, such as Watson (1993).
- Approaches which use the sampling variability of the actual data to provide a measure of distance between the model and the data. Among these are the GMM based approach of Eichenbaum and Christiano (1992) or Langot and Feve (1994), the indirect approach of Checchetti, Lam and Mark (1993) and the frequency domain approach of Diebold, Ohanian and Berkowitz (1998).
- Approaches which use the sampling variability of the simulated data to measure the discrepancy between the model and the data. Among these procedures we distinguish those which take the driving forces as stochastic and the parameters as given, such as Gregory and Smith (1991), Soderlin (1994) or Cogley and Nason (1994) and those who take both as random, such as Canova (1994), (1995) or Maffezzoli (2000).
- Approaches which use the sampling variability of both actual and simulated data to evaluate the fit. Again, we distinguish approaches which allow for variability in the parameters but not in the exogenous processes such as De Jong, Ingram and Whiteman (1995), (2000), Geweke (1999) and Schorfheide (2001) or allow both to vary, such as Canova and De Nicolo' (2003).

Roughly speaking, the first approach makes assumptions about the time series properties of L, given θ and ϵ_t; the second uses the sampling variability of S_y (and, in some cases, of θ) to evaluate the model; the third group of methods use sampling variability in ϵ_t in addition to either sampling variability or cross section variability of θ to evaluate the model; the final group of methods accounts for the variability in S_y, the cross sectional variability in θ and, in some cases, the variability in ϵ_t.

7.4.1 Watson's R^2

Standard statistical measures of fit use the size of the sampling errors to judge the coherence of a model to the data. That is, disregarding the approximation error, if ACF_y is the autocovariance function of the actual data and ACF_x is the autocovariance function of simulated data, standard measures examine whether $ACF_x = ACF_y$, given that differences between ACF_x and its estimated counterpart arises from sampling error. While this is a sensible procedure when the null represents the data, it is much less sensible when the model is false even under the null.

Rather than relying on the properties of the sampling error, Watson asks how much error should be added to x_t so that its autocovariance function equals the autocovariance function of y_t. The autocovariance function of this error is $ACF_v = ACF_y + ACF_x - ACF_{xy} - ACF_{yx}$, where ACF_{yx} is the cross-covariance function of x and y. Hence, to study the properties of ACF_v we need a sample from the joint distribution of (x_t, y_t) which is unavailable. Typically one of two assumptions is made (see e.g. Sargent (1989)): (i) $ACF_{xy} = ACF_x$, so that x_t and v_t are uncorrelated at all leads and lags. This yields a classical error-in-variables problem; (ii) $ACF_{xy} = ACF_y$, so that v_t is a signal extraction noise and y_t is the observable counterpart of x_t.

Example 7.8 *Let $y_t = x_t + v_t$ where $E(v_t v_{t-\tau}) = 0$, for $\tau \neq 0$ and equal to σ_v^2 when $\tau = 0$. Then $E(y_t x_{t-\tau}) = E(x_t x_{t-\tau})$ for all $\tau \neq 0$. Let now x_t be orthogonal to v_t, and let $x_t = \alpha y_t$. Then $E(y_t x_{t-\tau}) = \alpha E(y_t y_{t-\tau})$ for all $\tau \neq 0$.*

Clearly, which assumption is adopted depends on the way data is collected and expectations are formed. Here neither is very appealing since v_t is not a proxy nor a forecast error. Because any restriction used to identify ACF_{xy} is arbitrary, Watson chooses ACF_{xy} to minimize the variance of v_t, requiring ACF_x and ACF_y to be positive semidefinite. In other words, one selects ACF_{xy} so as to give the model the best chance to fit the data. The choice of ACF_{xy} depends on properties of the data and the dimension of x_t and y_t.

Example 7.9 *When x_t, y_t are serially uncorrelated scalars, the problem becomes $\min_{\sigma_{xy}} \sigma_v^2 = \sigma_x^2 + \sigma_y^2 - 2\sigma_{xy}$ subject to $\sigma_v^2 \geq 0$. The solution is $\sigma_{xy} = \sigma_x \sigma_y$. That is, selecting a minimum approximation error makes x_t and y_t perfectly correlated, and $x_t = \frac{\sigma_x}{\sigma_y} y_t$.*

The case of x_t, y_t serially uncorrelated, $m \times 1$ vectors, is analogous. The problem is now

$$\min_{\Sigma_{xy}} tr|\Sigma_v| = tr|\Sigma_x + \Sigma_y - \Sigma_{xy} - \Sigma_{yx}| \tag{7.14}$$

subject to $|\Sigma_v| \geq 0$ where $tr|\Sigma_v| = \sum_{i=1}^{m} \Sigma_{v_{ii}}$ is the trace of Σ_v. The solution is $\Sigma_{xy} = \mathcal{P}_x' V' \mathcal{P}_y$ where $\mathcal{P}_x$ and $\mathcal{P}_y$ are square roots of Σ_x and Σ_y, $V = \Omega\Lambda^{-1/2}\Omega'\mathcal{P}'$, Ω is a matrix of orthonormal eigenvectors and Λ is a diagonal matrix of eigenvalues of $\mathcal{P}'\mathcal{P}$ where $\mathcal{P} = \mathcal{P}_x\mathcal{P}_y$.

Exercise 7.10 *Describe how to compute predicted values of x_t given y_t when both are $m \times 1$ vectors. Argue that the joint covariance matrix of (x_t, y_t) is singular. Show how to modify (7.14) to minimize a weighted average of the diagonal elements of Σ_v.*

Typically in DSGE models, Σ_x is singular since the number of shocks is smaller than the number of endogenous variables. Let x_t, y_t be serially uncorrelated and let the m variables in x_t be driven by $m_1 \leq m$ shocks so that the rank of Σ_x is $m_1 \leq m$. Then, the above analysis applies to a $m_1 \times 1$ subvector of elements of x_t and y_t. Let S be a $m_1 \times m$ selection matrix such that $\mathsf{S}\Sigma_x\mathsf{S}'$ has rank m_1. Define $\tilde{x}_t = \mathsf{S}x_t$, $\tilde{y}_t = \mathsf{S}y_t$, $\tilde{\Sigma}_x = \mathsf{S}\Sigma_x\mathsf{S}'$, $\tilde{\Sigma}_y = \mathsf{S}\Sigma_y\mathsf{S}'$. Then $\tilde{v}_t = \tilde{x}_t - \tilde{y}_t$ is the error we wish to minimize. The solution is $\tilde{x}_t = \tilde{\mathcal{P}}_x' \tilde{V}' \tilde{\mathcal{P}}_y^{-1} \tilde{y}_t$ where $\tilde{\mathcal{P}}_x, \tilde{V}, \tilde{\mathcal{P}}_y^{-1}$ are the reduced rank analogs of $\mathcal{P}_x, V, \mathcal{P}_y^{-1}$.

Example 7.10 *Suppose $m = 2$ (say, output and consumption) and that $m_1 = 1$. Then we have three possible choice: we could minimize the variance of output, $\mathsf{S} = [1, 0]$, the variance of consumption, $\mathsf{S} = [0, 1]$, or a linear combination of the two, $\mathsf{S} = [\varrho, 1 - \varrho]$.*

Exercise 7.11 *Show that, because both Σ_x and $\mathsf{S}\Sigma_x\mathsf{S}'$ have rank m_1, it is possible to express x_t as a linear combination of $\tilde{x}_t$. Set $x_t = \mathcal{Q}\tilde{x}_t$ where $\mathcal{Q}$ is a $m \times m_1$ matrix. Conclude that $x_t = \mathcal{Q}\tilde{\Lambda}\mathsf{S}y_t$ and that it is optimal to set $\Sigma_{xy} = \mathcal{Q}\tilde{\Lambda}\mathsf{S}\Sigma_y$. Display the form of Lambda.*

When (x_t, y_t) are serially correlated vectors and $\text{rank}(\Sigma_x) = m_1 \leq m$, the same intuition applies. However, because of serial correlation, one wants to minimize the trace (weighted or unweighted) of the spectral density matrix. That is, we want to minimize $tr|W(\omega)\mathcal{S}_v(\omega)|$ where $W(\omega)$ is a matrix of weights for each frequency ω and $\mathcal{S}_{\tilde{v}(\omega)}$ is the spectral density matrix of v_t. When ω are Fourier frequencies $\mathcal{S}_{\tilde{v}(\omega)}$ is uncorrelated with $\mathcal{S}_{\tilde{v}(\omega)'}$, $\forall \omega \neq \omega'$. Hence the minimization problem can be solved frequency by frequency.

Exercise 7.12 *Show that the solution to the minimization problem when x_t, y_t are serially correlated and Σ_x is of rank m_1 is $ACF_{\tilde{x}\tilde{y}}(\omega) = \Lambda(\omega)ACF_{\tilde{y}}(e^{-i\omega})$ where $\Lambda(\omega)$ is the complex analog of Λ obtained in exercise 7.11. Show that $\tilde{x}_t$ is a function of leads and lags of $\tilde{y}_t$.*

Exercise 7.13 *Suppose that $x_t = \mathcal{Q}_1 v_t$ where x_t is a 2×1 vector, $\mathcal{Q}_1' = [1.0, 0.5]$ and $v_t \sim (0, 1)$ and let $y_t = \mathcal{Q}_2 e_t$, $\mathcal{Q}_2 = \begin{bmatrix} 1.0 & 0.3 \\ 0.2 & 1.0 \end{bmatrix}$ and e_t is a 2×1 vector of uncorrelated shocks with variances equal to 1 and 4. Show how to compute $ACF_{xy}(\omega)$ and the predicted value of x_t. Show both theoretical and numerical answers (the latter based on the estimation/ simulation of the relevant quantities).*

Once an expression for ACF_{xy} is obtained, it is easy to design R^2- type measures of fit. For example, we could use $\mathsf{S}_{1i}(\omega) = \frac{ACF_v(\omega)_{ii}}{ACF_y(\omega)_{ii}}$ or $\mathsf{S}_{2i}(\omega) = \frac{\int_{[\omega_1,\omega_2]} ACF_v(\omega)_{ii} d\omega}{\int_{[\omega_1,\omega_2]} ACF_y(\omega)_{ii} d\omega}$. S_{1i}

measures the variance of the i-th component of the error relative to the variance of the $i-th$ component of the data at frequency ω. Since this is analogous to $1-R^2$ in a regression, a plot of S_{1i} against ω visually provides a lower bound for the "distance" of the model from the data, frequency by frequency. S_{2i} may be useful to evaluate the model over a band of frequencies. Note that since v_t and x_t are serially correlated, both S_{1i} and S_{2i} can be greater than one.

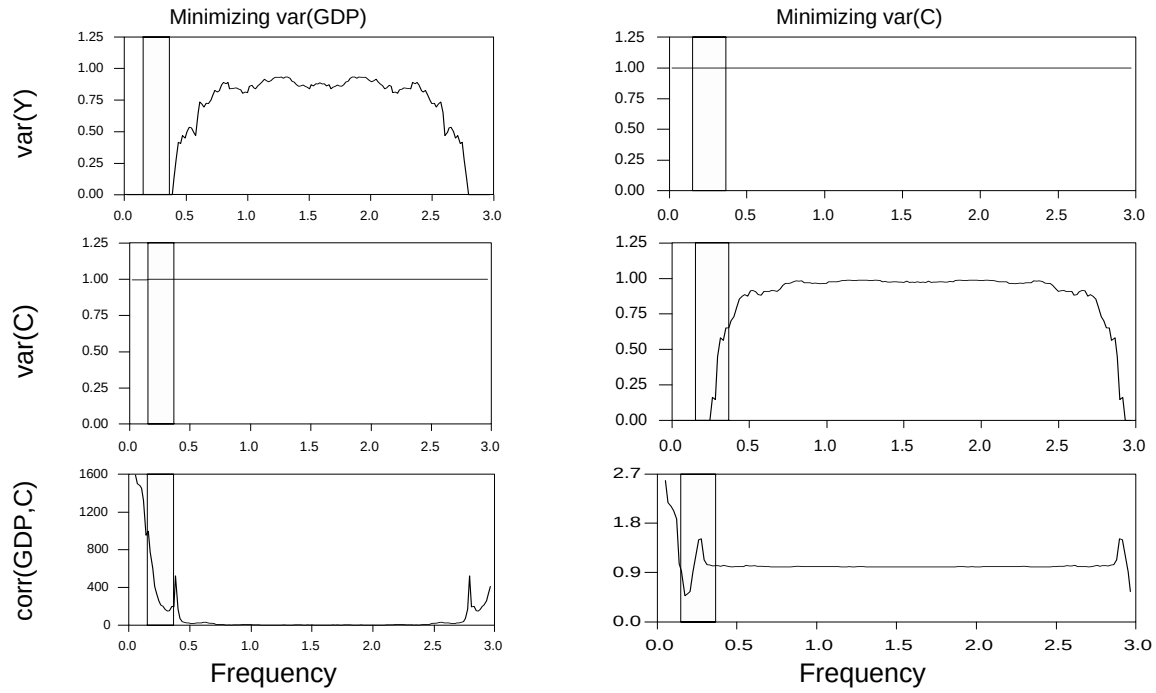


Figure 7.1: Watson's measure of fit

Exercise 7.14 *Show that linearly filtering x_t and y_t leaves S_{1i} unchanged but alters S_{2i}. (Hint: The weights depend on the frequency).*

Example 7.11 *We illustrate Watson's approach using simulated data from a version of the model described in example 7.3, where there are only technology shocks, the utility of the agents is $U(c_t, N_t) = \ln c_t + \vartheta_N(1-N_t)$ and $\eta = 0.64, \delta = 0.025, \beta = 0.99, (\frac{c}{GDP})^{ss} = 0.7, (\frac{K}{GDP})^{ss} = 2.5, \vartheta_N = 2.6, \bar{\zeta} = 0, \rho_\zeta = 0.95, \sigma_\zeta = 0.007$. Figure 7.1 presents S_{1i}, frequency by frequency, when we minimize the variance of output (first column) or the variance of consumption (second column). Here we care about the variability of output, the variability of consumption and their correlation. Actual data is linearly detrended. We need to choose*

one of the two variables because there is only one shock in the model. Shaded areas indicate business cycle frequencies. The model does not fit the data well: regardless of the minimization, $1-R^2$ is high for the variable whose variance is not minimized (roughly of the order of 0.9999) at business cycle frequencies. For the minimization variable, misspecification is noticeable at medium-high frequencies. Moreover, the correlation of two variables is poorly matched at low frequencies; misspecification declines with the frequency but at business cycle ones is still substantial.

Exercise 7.15 *(Wei) Consider two versions of a RBC model, one with habit persistence in leisure, and one with production externalities. In the first case the social planner maximizes $E_0\sum_t \beta^t[\ln(c_t)+\vartheta_N \ln(\gamma(\ell)(1-N_t))]$ subject to $c_t+K_{t+1}=K_t^{1-\eta}N_t^{\eta}\zeta_t$. In the second case she maximizes $E_0\sum_t \beta^t[\ln(c_t)+\vartheta_N \ln(1-N_t)]$ subject to $c_t+K_{t+1}=\bar{N}_t^{\aleph}K_t^{1-\eta}N_t^{\eta}\zeta_t$ where $\bar{N}_t$ is the average numbers of hour worked in the economy. In both cases we assume that $\ln\zeta_t=\bar{\zeta}+\rho_\zeta\ln\zeta_{t-1}+\epsilon_{1t}$. Suppose $\beta=0.99, \eta=0.64, \delta=0.025, \sigma_\zeta=0.007, \rho_\zeta=0.9$, ϑ_N is chosen so that $N^{ss}=0.20$, and $(\frac{c}{GDP})^{ss}=0.7$, $(\frac{K}{GDP})^{ss}=2.5$. Let $\gamma(\ell)=1+0.85\ell-0.3\ell^2$ and let $\aleph=0.45$. Compare the spectral densities of the two models for consumption, investment, output and hours to the spectra of US data and to the spectra of the model of example 7.11, linearly detrending the actual data and minimizing the variance of output or the arithmetic sum of variances of consumption, investment, output and hours. Which model improves over the basic one? In which dimensions?*

Exercise 7.16 *Consider a one sector monetary growth model with complete capital depreciation, inelastic labor supply, where the representative agent solves $\max_{c_t,K_{t+1}} E_0\sum_{t=1}^{\infty}\beta^t\ln c_t$ subject to $c_t+K_{t+1}\leq\zeta_t K_t^{1-\eta}N_t^{\eta}$ and $\ln\zeta_t=\bar{\zeta}+\rho_\zeta\ln\zeta_{t-1}+\epsilon_{1t}$, $\epsilon_{1t}\sim N(0,\sigma_\zeta^2)$. Assume a cash-in-advance constraint $M_t\leq p_t c_t$ and let $\ln M_t=\bar{M}+\rho_m\ln M_{t-1}+\epsilon_{3t}$, $\epsilon_{3t}\sim \mathsf{N}(0,\sigma_m^2)$.*
(i) Assume that the uncertainty is realized before decisions are taken at each t. Solve for the optimal path for (c_t, K_{t+1}, y_t, p_t) as a function of $(K_t,\ \zeta_t,\ M_t)$.
(ii) Assume $\eta=0.64,\ \beta=0.998,\ \rho_\zeta=0.90,\ \rho_m=0.8\ \sigma_\zeta=0.007,\ \sigma_m^2=0.01$. Apply Watson's approach to quarterly detrended money and price data in the US. Calculate $\mathsf{S}_{2i}(\omega)$ for $\omega\in[\frac{\pi}{16},\frac{\pi}{4}]$.

Two shortcoming of Watson's procedure should be noted. First, while there is some intuitive appeal in creating lower bounds statistics, it is not clear while one should concentrate only on the best possible fit. Canova, Finn and Pagan (1994) suggest to use both the best and the worst fit: if the range is narrow and $1-R^2$ of the worst outcome small, one can conclude that the model is satisfactory - a conclusion one can not reach using only the best fit. Second, the method does not provide information that may be useful in respecifying the model. R^2 could be low for a variety of reasons (the variance of the shocks in the data may be high; the dynamics of the model and of the data are different; the process for the states has a large AR coefficient). Clearly, it makes a lot of difference whether it is the first or the last of these causes that makes R^2 low.

7.4.2 Measure of fit based on simulation variability

A wide class of measures of fit can be obtained if a researcher is willing to randomize on the realization of the stochastic processes of the model. This approach was popularized, e.g., by Gregory and Smith (1991) and (1993). When the ϵ_t's are randomized, the distance between relevant functions of the model and of actual data can be evaluated using either asymptotic or probabilistic (Monte Carlo) criteria. Measures of fit of this type provide a sense of the "economic" distance between the model and the data, contrary, for example, to what one would obtain with likelihood ratio tests. Standard functions used for comparison include unconditional moments; spectral densities (see Soderlin (1994)), or semi-structural impulse responses (see Cogley and Nason (1994)). Note that spectra based comparisons do not require a parametric model for the actual data, but large simulated samples are needed to insure that the bias of the spectral estimates is small.

We illustrate how to validate a model with such a technique in the next example.

Example 7.12 *(Dunlop-Tarshis puzzle) Suppose we are concerned with the correlation between hours and labor productivity and suppose we are willing to draw replications for the time series of the random disturbances of the model. Then, there are three ways to check if the correlations produced by the model look like the actual ones. The first is as follows:*

Algorithm 7.2 1) *Draw* $(\{\epsilon_t\}_{t=1}^T)^l$. *Calculate* $ACRF_{N,w}(\tau)^l$, $l = 1, \ldots, L,\ \tau = 0, 1, 2, \ldots$

2) *Order simulations and construct percentiles.*

3) *Calculate the number of replications for which* $ACRF_{N,w}(\tau)^l$ *is less than the autocorrelation found in the actual data (separately for each* τ *or jointly); the decile of the simulated distribution whether the actual value lies or check whether the actual autocorrelation is inside a prespecified range of the simulated distribution (say, a 68% or a 95% interval).*

The output of algorithm 7.2 is what Gregory and Smith (1993) call the "size" of calibration tests. If a model is a poor approximation to the data, the simulated distribution of correlations will be far away from the distribution in the data and extreme statistics will be obtained (e.g. actual correlations are in the tails of the simulated distribution or the number of times $ACRF_{N,w}(\tau)^l$ *is less than the actual value is either zero or one).*

An alternative approach can be obtained using an asymptotic normal approximation for the distribution of correlations. For example, Anderson (1970) shows that $\widehat{ACRF}_{N,w}(\tau) \xrightarrow{D} \mathsf{N}(ACRF_{N,w}(\tau), \Sigma_{ACRF}(\tau))$ *where* $\Sigma_{ACRF}(\tau) = \frac{1}{2T}(1 - |ACRF_{N,w}(\tau)|)^2$. *Therefore, given one draw for* $\{\epsilon_t\}_{t=1}^T$, *and letting* $T \to \infty$; $\sqrt{T}\frac{ACRF_{N,w}(\tau) - \widehat{ACRF}_{N,w}(\tau)}{\sqrt{\Sigma_{ACRF}(\tau)}} \xrightarrow{D} \mathsf{N}(0,1)$, *each* τ. *Since little is know about the properties of correlation estimates when* T *is moderate or small, one may prefer a small sample version of this test in which case the next algorithm could be of use:*

Algorithm 7.3

1) *Draw replications for* $(\{\epsilon_t\}_{t=1}^{T})^l$ *and calculate* $ACRF_{N,w}(\tau)^l$ *,* $l = 1, \ldots L,\ \tau = 0, 1, 2, \ldots$.

2) *For each* l*, compare* $\sqrt{T}\frac{ACRF_{N,w}(\tau) - \widehat{ACRF}_{N,w}(\tau)}{\sqrt{\Sigma_{ACRF}(\tau)}}$ *to a* $\mathsf{N}(0,1)$*. Record either the p-value or construct a dummy variable which is one if the simulated distribution is statistically different from a* $\mathsf{N}(0,1)$ *at some confidence level and zero otherwise.*

3) *Construct the distribution of p-values or the percentage of times the model is rejected.*

We have examined whether the RBC model of example 7.3 driven by a government expenditure and a technology shocks can account for the correlation found in detrended US data for the sample 1964:1 to 2003:1. We use a version of the model with separable utility (a power specification for consumption and a linear specification for leisure). The parameters are $\beta = 0.99, \delta = 0.025, \vartheta = 0.5, \varphi = 2, N^{ss} = 0.2, \rho_\zeta = 0.9, \rho_g = 0.8, \sigma_\zeta = 0.007, \sigma_g = 0.01$.

	$\mathrm{Corr}(N_t, w_{t-1})$	$\mathrm{Corr}(N_t, w_t)$	$\mathrm{Corr}(N_t, w_{t+1})$
Size (% below actual)	0.40	0.27	0.32
Normality (% rejection)	0.59	0.72	0.66
Bands	[0.39, 0.65]	[0.45, 0.70]	[0.38, 0.64]
Actual correlations	0.517	0.522	0.488

Table 7.3: Cross correlation hours-wage

Table 7.3 reports the percentages of times the simulated correlation is below the actual one (row labelled "size"), the number of times the normality assumption is rejected (row labelled "normality") and the 68% band for simulated correlations together with the actual ones, for $\tau = -1, 0, 1$*. It is remarkable that actual correlations are inside the bands generated by the model at all three horizons. Also while the model has a tendency to produce correlations which exceed those found in the actual data, the results are reasonably good. This is not completely surprising. As suggested in exercise 1.15 in chapter 2, the presence of demand shifters can reduce to realistic levels the almost perfect correlation between hours and real wages produced by technology shocks.*

Exercise 7.17 *(Adelmann test) Consider two versions of a RBC model, one with habit persistence in consumption and one with one-period labor contracts. Assume that there are only productivity shocks, that the solutions are obtained log-linearizing the optimality conditions and that the productivity process is parametrized so that it reproduces the first two moments of actual Solow residuals for the US economy.*
(i) Appropriately select the remaining parameters of both models.
(ii) Construct probabilities of turning points in output, defining a recession at t *if* $gdp_{t-2} > gdp_{t-1} > gdp_t < gdp_{t+1} < gdp_{t+2}$ *and an expansion at* t *if* $gdp_{t-2} < gdp_{t-1} < gdp_t > gdp_{t+1} > gdp_{t+2}$ *(Hint: draw sequences for the exogenous disturbances and count the number of times that at each date recession and expansion events are encountered).*
(iii) Design a probabilistic statistic to assess which model fits the NBER chronology better.

Exercise 7.18 *(Money-inflation relationship) Consider a working capital economy (like the one in exercise 1.14 of chapter 2) and a sticky price economy (like the one in example 1.5 of chapter 2). Suppose we want to find out which model fits the actual cross correlation function of money and inflation found in Euro area data better. Assume that there are two shocks in each model (a technology and a monetary one), both of which are AR(1); that in both models output requires capital and labor; that there is no habit persistence in consumption, that there are quadratic costs to adjusting capital of the form* $\frac{b}{2}(\frac{K_{t+1}}{K_t}-1)^2K_t$ *where* $b \geq 0$ *and that monetary policy is conducted according to a rule of the form* $i_t = i_{t-1}^{a_0} gdp_t^{a_1} \pi_t^{a_2} \epsilon_{3t}$ *where* ϵ_{3t} *is a monetary policy shock and* i_t *the nominal interest rate. Log-linearize both models around the steady state, appropriately select the parameters and construct probabilistic measures of fit, randomizing on the stochastic processes for monetary and technology shocks.*

Instead of measuring the distance between moments, one could measure distance between structural impulse responses, where structural shocks are obtained with one of the approaches described in chapter 4. A statistic to compare impulse responses is:

$$\mathsf{S}(\tau) = [IRF(\tau) - IRF^A(\tau)]\Sigma(\tau)^{-1}[IRF(\tau) - IRF^A(\tau)]' \tag{7.15}$$

where $\tau = 1, 2, \ldots$ refers to the horizon, $IRF(\tau)$ is the mean response of the model (across replications) at horizon τ, $IRF^A(\tau)$ is the actual response at horizon τ and $\Sigma(\tau) = \frac{1}{L}\sum_{l=1}^{L}[IRF(l,\tau) - IRF(\tau)^A][IRF(l,\tau) - IRF(\tau)^A]'$. Asymptotically, $T \times \mathsf{S}(\tau) \sim \chi^2(1)$. For the sample sizes used in macroeconomics, a small sample version of this statistic is probably more useful. Hence,

$$\mathsf{S}(\tau, l) = [IRF(\tau, l) - IRF^A(\tau)]\Sigma(\tau)^{-1}[IRF(\tau, l) - IRF^A(\tau)]' \tag{7.16}$$

could be used where $IRF(\tau, \ell)$ is the response of the model for horizon τ and replication ℓ. As suggested in algorithm 7.3, $\mathsf{S}(\tau, l)$ can be computed at each l, using the simulated realization of impulse responses. Then, the empirical distribution of $\mathsf{S}(\tau, l)$ can be constructed and the rejection frequency computed. Since $\Sigma(\tau)$ is correlated across τ, it is necessary to eliminate the correlation if a joint comparison at more than one τ needs to be made (see e.g. chapter 4).

Exercise 7.19 *(Long run neutrality) Continuing with exercise 7.18, in both models the long run correlation between money growth and inflation is one. Using algorithm 7.3 or a statistic similar to (7.16) provide a probabilistic assessment of whether the actual long run correlation of money growth and inflation could have been generated by these two models.*

Exercise 7.20 *(Output and prices) Continuing with the model economy described in exercise 7.16, run a VAR on simulated price and income data, and identify structural shocks using the assumption that, on impact, one of the shocks has no effect on prices.*
(i) Repeat the identification exercise for actual data.
(ii) Calculate $\mathsf{S}(\tau, l)$, $\tau = 1, 4, 8$. *Tabulate the rejection frequencies and interpret the results.*

The above setup can be easily modified to account for parameter uncertainty. To do this we only need to change the first step of algorithms 7.2 and 7.3, randomizing both e_t and θ. The empirical distribution of relevant statistics can then be constructed and from there one can compute size tests or percentiles rejection rates, i.e. calculate in what percentage of the model distribution the actual statistics lie. When parameter uncertainty is "objective", as in Canova (1994)-(1995), the extension is straightforward: we simply draw from the assumed joint empirical distribution for the parameters. When parameters uncertainty is data-based, the techniques described in the next subsection could be used.

7.4.3 Measures of fit based on sampling variability

If one allows estimation variability in the parameters, or if one is willing to accept the idea that stylized facts are measured with error, model evaluation can be conducted using a metric which exploits sampling, as opposed to simulation, variability.

When parameters are random the procedure typically used resembles a J-test (see chapter 5). However, here simulated moments are random - because of parameter uncertainty - while moments of the actual data are assumed to be measured without error. Hence, data moments play the role of g_∞ and simulated moments the role of g_T in the GMM setup.

Suppose θ_T solves $\frac{1}{T}\sum_t g_1(\epsilon_t, \theta)$, let $g_2(y_t)$ be another vector of moments of actual data, $g_2(\epsilon_t, \theta_T)$ the same vector of moments obtained from simulated data and Σ_{g_2} the covariance matrix of $g_2(\epsilon_t, \theta_T)$. Then, as $T \to \infty$, $T \times [g_2(y_t) - g_2(\epsilon_t, \theta_T]' \Sigma_{g_2}^{-1} [g_2(y_t) - g_2(\epsilon_t, \theta_T)] \sim \chi^2(dim(g_2))$. Note that $\Sigma_{g_2} = \frac{\partial g_2}{\partial \theta} \Sigma_\theta \frac{\partial g_2}{\partial \theta}'$ where Σ_θ is the covariance matrix of θ_T.

Exercise 7.21 *Assume that $g_2(y_t)$ is measured with error. Show how to modify the distance statistic and its asymptotic distribution to take this into account.*

Two points need to be stressed. First, the method is closely related to those discussed in chapter 5. Therefore, standard conditions on y_t and on the g functions are required for the statistics to have asymptotic validity. Note also that here estimation and testing are conducted sequentially, as opposed to simultaneously, and that θ_T is obtained from just identified conditions. Second, a J-test is valid under the null the model is the true DGP in the dimensions represented by g_2. That is, the model needs to be correct at least in the g_2 dimensions for the validation results to have meaningful interpretations.

An alternative approach, not requiring the assumption that the model is true, was suggested by Diebold, Ohanian and Berkowitz (DOB) (1998). The method is close in spirit to Watson's but uses the sampling variability of the actual data to construct a finite sample diagnostic of fit.

Let $\mathcal{S}_y(\omega)$ be the spectrum of the actual data and $\widehat{\mathcal{S}}_y(\omega)$ an estimate of $\mathcal{S}_y(\omega)$. When y_t is univariate and T large, $\frac{2\hat{\mathcal{S}}_y(\omega)}{\mathcal{S}_y(\omega)} \sim \chi^2(2)$ for $\omega \neq 0, \pi$. For the sample sizes typically used in macroeconomics, asymptotic approximations are probably inappropriate and DOB suggest two bootstrap methods to construct small sample confidence intervals for the spectrum of y_t. The methods differ in the way replications are constructed: in the first the asymptotic

distribution of the prediction error is the resampling distribution; in the second it is its empirical distribution. Let $\tilde{\mathcal{P}}'\tilde{\mathcal{P}} = ACF_y(\tau) \equiv cov(y_t y_{t-\tau})$ and let $\bar{y}$ be the sample mean of y_t. The methods are summarized in the following algorithm:

Algorithm 7.4

1) *Draw* v^l *from a* $\mathsf{N}(0, I_T)$ *or from the empirical distribution of* $v_t = \tilde{\mathcal{P}}^{-1}(y_t - \bar{y})$.

2) *Construct* $y_t^l = \bar{y} + \tilde{\mathcal{P}} v_t^l$ *and* $ACF^l(\tau)$, $l = 1, \ldots, L$.

3) *Compute* $\hat{\mathcal{S}}_y^l(\omega) = \sum_\tau \mathcal{K}(\tau) ACF_y^l(\tau) e^{-i\omega\tau}$, *where* $\mathcal{K}(\tau)$ *is a kernel.*

4) *Order* $\hat{\mathcal{S}}_y^l(\omega)$, *each* ω; *construct percentiles and extract confidence intervals.*

Note that bootstrap procedures are valid under homoschedasticity of v_t. Therefore, if heteroschedasticity is suspected, data needs to be transformed before algorithm 7.4 is used.

Exercise 7.22 *Describe a bootstrap approach to compute small sample confidence intervals for* $\mathcal{S}_y(\omega)$, *drawing* $\frac{2\hat{\mathcal{S}}_y(\omega)}{\mathcal{S}_y(\omega)}$ *from a* $\chi^2(2)$ *or from its empirical distribution.*

The multivariate analogs of these estimators are straightforward.

Exercise 7.23 *Describe how to implement the parametric bootstrap algorithm 7.4 and the non-parametric bootstrap algorithm of exercise 7.22 in a multivariate setting.*

Bootstrap distributions are valid frequency by frequency. However, evaluation is often performed over a band of frequencies. The results obtained by connecting the p-values, frequency by frequency, are incorrect since a set of n $(1-\varrho)\%$ confidence intervals constructed for each frequency will not achieve a $(1-\varrho)\%$ joint coverage. Rather, the actual confidence level will be closer to $(1-\varrho)^n\%$ if the pointwise intervals are independent. Hence, when interest centers in a band of n frequencies, a more appropriate approximation is obtained choosing a $(1-\varrho/n)\%$ coverage for each spectral ordinate since the resulting tunnel has coverage of, at least, $(1-\varrho)\%$.

When the parameters and the stochastic processes of a model are fixed at some $\hat{\theta}$ and $\hat{\epsilon}_t$, the spectrum of simulated data can be constructed to any degree of accuracy either by simulating a very long time series or by replicating many times a short time series using the distribution of $\hat{\epsilon}_t$ and invoking ergodicity. Let $\mathcal{S}_x(\omega, \hat{\theta}, \hat{\epsilon}_t)$ be the spectrum of the model. A measure of fit is:

$$\mathsf{L}(\hat{\theta}, \hat{\epsilon}_t) = \int_{\omega_1}^{\omega_2} \mathsf{L}^*(\mathcal{S}_y(\omega), \mathcal{S}_x(\omega, \hat{\theta}, \hat{\epsilon}_t)) W(\omega) d\omega \tag{7.17}$$

where $W(\omega)$ is a set of weights and L^* is a function measuring the distance between the spectrum of actual and simulated data at frequency ω.

Exercise 7.24 *Show the form of* $\mathsf{L}(\hat{\theta}, \hat{\epsilon}_t)$ *when* L^* *is quadratic and when we are interested in comparing model and data at business cycle frequencies only. Show that if* $\mathsf{L}^* = \frac{\mathcal{S}_x(\omega, \hat{\theta}, \hat{\epsilon}_t)}{\mathcal{S}_y(\omega)}$ *detrending is irrelevant in judging the closeness of the model to the data.*

It easy to include parameter uncertainty into the evaluation criteria. In fact, one advantage of evaluating the fit as in (7.17) is that $\mathsf{L}(\hat{\theta}, \hat{\epsilon}_t)$ can also be used for estimation purposes. For example, $\tilde{\theta} = argmin_{\theta}\mathsf{L}(\theta, \hat{\epsilon}_t)$ is a minimum distance type estimator of θ. Recall that GMM is an estimator of this form, while maximum likelihood has asymptotically this form (here we would set $W(\omega_j) = 1,\ \forall\omega_j$ and $\mathsf{L}(\theta, \hat{\epsilon}_t) = -0.5\sum_j \ln \mathcal{S}_x(\omega_j, \theta, \hat{\epsilon}_t) - 0.5\sum_j \frac{\mathcal{S}_y(\omega_j)}{S_x(\omega_j, \theta, \hat{\epsilon}_t)}$). When this is done the assumption that the model is correct, at least in some dimensions, is necessary for the estimation-evaluation process to make sense.

It is also easy to combine the algorithm 7.4 with estimation: the first three steps of the procedure are identical and we only need to estimate $\tilde{\theta}^l$ for each draw, $l = 1, \dots L$. From the distribution of $\tilde{\theta}^l$ we can construct point estimates, confidence intervals, etc.

Exercise 7.25 *Suppose* L^* *is quadratic. Show how the variability of* θ *and of* $\mathcal{S}_x(\omega)$ *affect estimates of* $\mathsf{L}(\theta, \epsilon_t)$.

Example 7.13 *(Band spectrum regression) Suppose* ω_j *are Fourier frequencies and concentrate attention at business cycle frequencies (i.e.* $[\frac{\pi}{16}, \frac{\pi}{4}]$*). Suppose that* L^* *is quadratic,* $W(\omega) = 1,\ \forall\omega$ *and* $\mathsf{L}(\theta, \hat{\epsilon}_t) = \sum_j (\mathcal{S}_y(\omega_j) - \mathcal{S}_x(\omega_j, \theta, \hat{\epsilon}_t))^2$*. When* $\mathcal{S}_x(\omega, \theta, \hat{\epsilon}_t)$ *is orthogonal to* $\mathcal{S}_y(\omega_j) - \mathcal{S}_x(\omega_j, \theta, \hat{\epsilon}_t)$*, minimization of* $\mathsf{L}(\theta, \hat{\epsilon}_t)$ *produces band spectrum regression estimates (see, Engle (1974)).*

Exercise 7.26 *For the economy of exercise 7.16 calculate the spectrum of prices and income, appropriately selecting the parameters. Calculate the spectrum of prices and income using US data and measure the distance of the two using a parametric bootstrap algorithm.*
(i) Calculate $\mathsf{L}(\hat{\theta}, \hat{\epsilon}_t)$ *assuming that* L^* *is quadratic, equally weighting deviations of prices and income from their actual counterpart at business cycle frequencies.*
(ii) Repeat the calculation when $\mathsf{L}(\theta, \hat{\epsilon}_t) = \mathcal{I}_{[\mathcal{S}_y(\omega) \geq \mathcal{S}_x(\omega)]}$ *and* $\mathcal{I}$ *an indicator function. Find the* $\hat{\theta}$ *which minimizes this quantity.*

Exercise 7.27 *Design an asymptotic criteria to compare the spectra of a model and of the data. Describe how to use it to compare alternative calibrated models.*

Example 7.14 *Continuing with the economy of example 7.11, we compute joint 68% tunnels for the spectra of consumption and output and for the coherence between the two variables using a parametric bootstrap approach. Figure 7.2 shows the tunnels together with the log spectra and coherence produced by the model (business cycle frequencies are highlighted). Clearly, consumption and output variability in the model at business cycle frequencies are lower in the data, while the coherence is, roughly, of the same magnitude.*

7.4.4 Measures of fit based on sampling and simulation variability

There is no reason to confine attention to either sampling or simulation uncertainty. The outcomes of models are uncertain because parameters (and forcing variables) are unknown;

Figure 7.2: Spectra and Coherences

statistics of the data are uncertain because of sampling variability. Therefore, it makes sense to design a metric which takes both types of uncertainty into account.

Canova and De Nicolo' (2003), for example, construct bootstrap distributions for statistics of actual data and simulated distributions for the statistics produced by the model (allowing for uncertainty in both the parameters and in the stochastic processes) and measure the quality of the approximation by examining the degree of overlap between the two distributions - a large overlap for different contour probabilities is considered a good sign. Actual and simulated data are treated symmetrically: one can either ask if the actual data could have been generated by the model or, viceversa, if the simulated data are consistent with the empirical distribution of the actual data. Roughly speaking this is equivalent to the process of switching the null and the alternative in hypothesis testing.

Example 7.15 *Continuing with example 7.12, we ask how much overlap there is between the distributions of the contemporaneous correlation of hours and real wages in the data and in the model when uncertainty in both parameters and stochastic processes is taken into account. Figure 7.3 reports the two distributions: there appears to be some overlap but the simulated distribution is much more spread out than the actual one: in fact, the small sample 95% interval of actual correlation (0.44,0.52) is inside the 68% interval for the model correlation (0.35, 0.75). Viceversa, only the central 25% of the mass of the simulated distribution of the model correlation is inside the 68% interval of the small sample distribution of the actual one.*

Figure 7.3: Distributions of Hours and Labor Productivity Correlation

DeJong, Ingram and Whiteman (1996) also take the view that uncertainty characterizes both models and data. However, distributions characterizing the uncertainty in the parameters of the model and in the coefficients of the parametric representation of the data are "subjective". Suppose we represent a $m \times 1$ vector of actual time series with a VAR:

$$y_t = A(\ell)y_{t-1} + e_t \quad e_t \sim \mathsf{N}(0, \Sigma_e) \tag{7.18}$$

Let $\alpha = \{vec(A_1), vec(A_2), \ldots, vech(\Sigma_y)\}$ where $vec(.)(vech(.))$ are the columnwise vectorizations of rectangular (symmetric) matrices and let $\Sigma_y = (I - A(\ell)\ell)^{-1}\Sigma_e(I - A(\ell)\ell)^{-1'}$. When y_t is stationary, and given α, the second moments of y_t can be obtained as $\Sigma_Y = \mathsf{A}\Sigma_Y\mathsf{A}' + \Sigma_E$ where A is the companion matrix of $A(\ell)$ and $ACF_Y(\tau) = \mathsf{A}^\tau\Sigma_Y$. Let $g(\alpha) = \prod_i g(\alpha_i),\ i = 1, 2, \ldots$ be a prior density for α. We let $g(\alpha)$ to be noninformative and require that $g(\alpha_i)$ is such that y_t is stationary, i.e. $g(\alpha) \propto |\Sigma_E|^{\frac{m+1}{2}} \times \mathcal{I}_{[stationary]}$ where $\mathcal{I}_{[stationary]}$ is an indicator function. As discussed in details in chapter 10, a uninformative prior for α and a normal likelihood for y_t generate a Normal-Wishart posterior distribution for α. Then, the posterior distribution for Σ_Y, ACF_Y can be obtained by simulation, drawing α from such a distribution, computing Σ_Y and ACF_Y for each draw and collecting relevant percentiles.

Let θ be the vector of parameters of the model. The outcomes of the model can be described by a density $f(x_t|\theta)$. Let $g(\theta)$ be a prior density for θ. Then $f(x_t) = \int f(x_t|\theta)g(\theta)d\theta$ characterizes the realizations of the model, once parameter uncertainty is averaged out. Using draws from $g(\theta)$ and the solution to the model $f(x_t|\theta)$, a simulation-based distribution of Σ_x and ACF_x, accounting for parameter uncertainty, can be produced.

Given a functional form for $g(\theta)$ (say, a joint normal distribution) one can vary its dispersion to see if the degree of overlap between the distribution of Σ_x and of Σ_y (or ACF_x and ACF_y) increases. Since there is disagreement in the profession on the spread of $g(\theta)$, such an exercise may help to understand whether uncertainty in the selection of the parameters θ can mitigate differences between the model and the data.

Let $g(\mathsf{S}_{y,i})$ be the data based distribution of the component i of S_y and let $g_j(\mathsf{S}_{x,i})$ be the model based distribution for $\mathsf{S}_{x,i}$ using specification j of the prior. One way to measure the degree of overlap between $g(\mathsf{S}_{y,i})$ and $g(\mathsf{S}_{x,i})$ is the following Confidence Interval Criterion (CIC)

$$CIC_{ij} = \frac{1}{1-\varrho}\int_{\varrho/2}^{1-\varrho/2} g_j(\mathsf{S}_{x,i})d\mathsf{S}_{x,i} \tag{7.19}$$

where $1-\varrho = \int_{\varrho/2}^{1-\varrho/2} g(\mathsf{S}_{y,i})d\mathsf{S}_{y,i}$. Note that $0 \leq CIC_{ij} \leq \frac{1}{1-\varrho}$. When CIC_{ij} is low, the fit is poor i.e. either the overlap is small or $g_j(\mathsf{S}_{x,i})$ is very diffuse; values close to $\frac{1}{1-\varrho}$ indicate that the two distributions overlap substantially and values greater than one that $g(\mathsf{S}_{y,i})$ is diffused relative to $g_j(\mathsf{S}_{x,i})$. To distinguish among the two interpretations when CIC_{ij} is low, one can supplement (7.19) with another measure, analogous to a t-statistic for the mean of $g_j(\mathsf{S}_{x,i})$ in the $g(\mathsf{S}_{y,i})$ distribution, i.e., $\frac{Eg_j(\mathsf{S}_{x,i})-Eg(\mathsf{S}_{y,i})}{\sqrt{var(g(\mathsf{S}_{y,i}))}}$. Large values of this statistic relative to a $\mathsf{N}(0,1)$ indicate differences in the location of $g_j(\mathsf{S}_{x,i})$ and $g(\mathsf{S}_{y,i})$.

While so far we have kept ϱ fixed, it is probably a good idea to vary it, given j, since large differences produced by different values of ϱ provide information on how and where the two distributions overlap.

Exercise 7.28 *Consider the economy analyzed in exercise 7.4.*
(i) Using US data on equity returns and the risk free rate calculate the equity premium as $EP_t = (R_t^e - R_t^f)$ *and the mean and the autocovariance of* R_t^f *and* EP_t *at lags -1, 0, 1.*
(ii) Set $n = 2, gy_1 = 0.9873, gy_2 = 1.0177, p_1 = 0.2, p_2 = 0.8, \beta = 0.99, \rho_y = 0.8, \varphi = 2$. *Simulate asset price data from the model and compute* $\mathsf{S}_x = [E[(R_t^f), (R_t^e)];\ var[(R_t^f), (R_t^e)]$.
(iii) Consider the second moments of the two variables and assume that, as in Watson's approach, you want to minimize the variance of the risk free rate. Provide individual and joint measures of fit for the risk free rate and the equity premium at all frequencies.
(iv) Examine the fit of the model for the mean of the equity premium and of the risk free rate, using a metric based on the sampling variability of simulated data.
(v) Describe a bootstrap algorithm to calculate the small sample distribution of the variance of the two variables. Using a quadratic loss function, find the φ *which produces the best fit of the model to the data using Diebold, Ohanian and Berkowitz approach.*
(vi) Assume $\varphi \sim \mathsf{N}(2.0, 0.1),\ \beta \sim \mathsf{N}(0.96, 0.01), \rho_y \sim \mathsf{N}(0.8, 0.05)$ *and still let* $gy_1 = 0.9873, gy_2 = 1.0177, p_1 = 0.2, p_2 = 0.8$ *Draw 100 values for these parameters, compute* S_x *for each draw and calculate (plot) the joint empirical distribution. Repeat the exercise assuming* $\varphi \sim \mathsf{N}(2.0, 0.2)$ *and still assuming* $\beta \sim \mathsf{N}(0.96, 0.02), \rho_y \sim \mathsf{N}(0.8, 0.05)$, $gy_1 = 0.9873, gy_2 = 1.0177, p_1 = 0.2, p_2 = 0.8$.

(vii) Run a bivariate VAR using a measure of real risk free rate of interest and of the equity premium. Draw 100 parameters from a Normal-Wishart distribution for the VAR coefficients, compute S_y *from each draw and plot the joint distribution.*
(viii) Show the degree of overlap between the distributions you have computed in (vi) and in (vii) letting $\varrho = 0.01, 0.10$*. Calculate CIC in the two cases.*
(ix) In what percentile of the distribution of S_y *lies the value of* S_x *computed in (ii)?*

One can notice that there is still some asymmetry in the procedure: we compare the predictive density of the model $f(x_t)$ and the posterior distribution of the data. In principle, one would like to use the posterior distribution of both the model and the data. However, to construct posterior distributions for the parameters, we need Monte Carlo Markov chain methods. We defer the discussion of these methods to chapter 9.

Hansen and Heckman (1996) criticize users of the computational experiments on the ground that they rarely perform out-of-sample forecasting comparisons between the model and simple time series specifications. The setup used in this chapter allows for this type of exercises. Such a comparison is useful in two senses. First, since DSGE models are restricted VAR, a comparison with unrestricted VARs help us to gauge the validity of the restrictions. Second, if a DSGE is not too bad in forecasting relative to time series models, policymakers may have an incentive to take the reported measurement more seriously.

Example 7.16 *(Using RBC model to forecast output)*
We take once again the specification used in example 7.3 and use it to forecast output. Conditional on the parameters, the model produces forecasts for $gdp_t(\tau), \tau = 1, 2, \ldots$ *at every* t *via the recursive formula* $\mathsf{S}y_{t+\tau} = A_0^G A_1 \mathsf{S}y_{t+\tau-1}$*, where* S *is a selection matrix which extracts* gdp_t *from the vector* y_t *and* A^G *is the generalized inverse of* A*. We compare the forecasts of the model to those produced by a naive random walk model, i.e.* $gdp_t(\tau) = gdp_t \forall \tau$ *by computing the ratio of the MSE of the two models. For the sample 1990:1-2001:3 such a statistic at one-step horizon is 1.13. At four steps however its value drops to 0.97, so that the restrictions imposed by the model help at somewhat longer horizons. If we randomize on the parameters of the model we can construct small sample distributions for the forecasts and for the ratio of MSEs. In this case the ratio of the two MSEs is greater than one in more than 70% of the cases at the one-step horizon, but less than one in 58% of the cases at the four step horizon.*

The setup also allows for other types of forecasting comparisons as it is explained next.

Example 7.17 *Consider a bivariate VAR model with money and output and assume a flat prior on* α*. From the posterior distribution of* α *one can draw realizations* α^l *and use (7.18) to forecast recursively out-of-sample. Once a prior for the deep parameters* θ *of the model is available, a representation like (7.13) allows us to compute forecasts* $\mathsf{S}y^l_{t+\tau}$*,* $\tau = 1, 2, \ldots$ *for every draw* θ^l*. Then to compare the two sets of forecasts at each* τ*, one can compute the number of times the MSE of the model is lower than the MSE of the VAR or relate* $T \times (MSE(\theta) - MSE(\alpha))\Sigma^{-1}(MSE(\theta) - MSE(\alpha))'$ *to a* $\chi^2(1)$ *distribution where* Σ *measures the dispersion of* $MSE(\theta, l)$ *and* $MSE(\alpha, l)$ *from their mean.*

It is important to stress that when one uses a DSGE model to forecast out-of-sample, one does not necessarily subscribe to the idea the model is the DGP. In fact, comparisons based on MSEs do not require such an assumption: forecasts could be acceptable even when estimates are biased if the variance of the forecasts is small. Similarly, graphical analyses, the examination of historical events or turning points do not require such an assumption.

We have repeatedly emphasized that calibrators and standard econometricians take a different point of view regarding the nature of an economic model. For a standard econometrician, the distribution of outcomes of the model is the probability density function of the data which can be used as a likelihood function to conduct inference. Calibrators which choose parameters using GMM or similar techniques implicitly assume that the model describes only selected features (moments) of the real data. Call these features S_y. Since the relationship between S_y and the outcomes of the model is analogous to the one assumed by traditional econometricians, the same evaluation techniques can be used. However, in constructing the mapping between S_x and S_y, it is easy to fall into logical inconsistencies. An example can clarify why this is the case.

Example 7.18 *Suppose one is interested in constructing a model to explain the average returns from holding financial assets. Let y_{it} be the return of asset i produced by the model, let $\bar{y}_i$ be its first moment and assume that returns are iid with constant variance. Then, the sampling distribution of the $\bar{y}_i$'s depends on the population variance of returns since, $var(\bar{y}_i) = \frac{var(y_{it})}{T}$. Hence, features of the data that the model wants to explain (first moment of returns) may depend on features that the model is not intended to explain (second moments of the returns).*

To avoid inconsistencies, Geweke (1999) suggests to view a DSGE model as a representation for the *population moments of observable functions* of the data - not of their *sample* counterparts. This setup is advantageous because comparisons across models do not require the likelihood function of the data. However, since DSGE models have no interpretation for the observables, it is necessary to bridge population and sample statistics.

Formally, let S_y be a vector of functions of a subset of the data, let $\mathcal{M}_1$ and $\mathcal{M}_2$ be two different model specifications with parameters θ_1 and θ_2 respectively; let $\mathsf{S}_{\infty,1} = E[\mathsf{S}_y|\theta_1, \mathcal{M}_1]$ and $\mathsf{S}_{\infty,2} = E[\mathsf{S}_y|\theta_2, \mathcal{M}_2]$ be the population functions the models produce and let $f(\mathsf{S}_{\infty,1}|\mathcal{M}_1)$ and $f(\mathsf{S}_{\infty,2}|\mathcal{M}_2)$ be the densities for S_∞ induced by the two models. Let the prior on the parameters be $g(\theta_1)$ and $g(\theta_2)$ and let $\mathcal{M}_3$ be a time series model which allows to compute the posterior distribution of S_∞, denoted by $g(\mathsf{S}_{\infty,3}|y_t, \mathcal{M}_3)$, given the observables y_t.

Assume that $f(y_t|\mathsf{S}_{\infty,3}, \mathcal{M}_1, \mathcal{M}_3) = f(y_t|\mathsf{S}_{\infty,3}, \mathcal{M}_2, \mathcal{M}_3) = f(y_t|\mathsf{S}_{\infty,3}, \mathcal{M}_3)$ and that $g(\mathsf{S}_{\infty,1}|\mathcal{M}_1, \mathcal{M}_3) = f(\mathsf{S}_{\infty,1}|\mathcal{M}_1)$, $g(\mathsf{S}_{\infty,2}|\mathcal{M}_2, \mathcal{M}_3) = f(\mathsf{S}_{\infty,2}|\mathcal{M}_2)$. Intuitively, we require that knowledge of the two models carries no information for y_t (they are assumed to describe S_∞) and that $\mathcal{M}_3$ has nothing to say about S_∞, either absolutely, or relative to $\mathcal{M}_1$ and $\mathcal{M}_2$.

Exercise 7.29 *Show that if $g(\mathsf{S}_{\infty,3}|\ \mathcal{M}_3)$ is a constant and $g(\mathsf{S}_{\infty,1}|\ \mathcal{M}_1,\ \mathcal{M}_3)\ =$*

$f(\mathsf{S}_{\infty,1}|\mathcal{M}_1)$, *the posterior of model* $\mathcal{M}_1$, *given* y_t *and the empirical model* $\mathcal{M}_3$, *is:*

$$g(\mathcal{M}_1|y_t,\mathcal{M}_3) \propto g(\mathcal{M}_1|\mathcal{M}_3)\int f(\mathsf{S}_{\infty,1}|\mathcal{M}_1)g(\mathsf{S}_{\infty,1}|y_t,\mathcal{M}_3)d\mathsf{S}_{\infty,1} \tag{7.20}$$

so that a posterior odds ratio for the two models is

$$\frac{g(\mathcal{M}_1|y_t,\mathcal{M}_3)}{g(\mathcal{M}_2|y_t,\mathcal{M}_3)} = \frac{g(\mathcal{M}_1|\mathcal{M}_3)}{g(\mathcal{M}_2|\mathcal{M}_3)}\frac{\int f(\mathsf{S}_{\infty,1}|\mathcal{M}_1)g(\mathsf{S}_{\infty,1}|y_t,\mathcal{M}_3)d\mathsf{S}_{\infty,1}}{\int f(\mathsf{S}_{\infty,2}|\mathcal{M}_2)g(\mathsf{S}_{\infty,2}|y_t,\mathcal{M}_3)d\mathsf{S}_{\infty,2}} \tag{7.21}$$

Equations (7.20)-(7.21) show two important facts. First, the posterior distribution of a model is proportional to the product of the density of the model for S_∞, and its posterior obtained using the empirical model $\mathcal{M}_3$ and the data y_t, with a factor of proportionality depending on the prior of the model. Second, the posterior odds of model $\mathcal{M}_1$ relative to model $\mathcal{M}_2$ depend on the degree of overlap of $f(\mathsf{S}_\infty|.)$ with the posterior distribution of S_∞, given $\mathcal{M}_3$. Hence, model $\mathcal{M}_1$ is preferable to model $\mathcal{M}_2$ if the overlap of the distribution of S_∞ produced by $\mathcal{M}_1$ with its posterior distribution computed using $\mathcal{M}_3$ is higher than the overlap of the distribution of S_∞ produced by $\mathcal{M}_2$ with its posterior computed using $\mathcal{M}_3$. The term $\frac{g(\mathcal{M}_1|\mathcal{M}_3)}{g(\mathcal{M}_2|\mathcal{M}_3)}$ represents the prior odds of the two models, given $\mathcal{M}_3$. Since y_t could be a vector, (7.21) extends the univariate confidence interval criteria used by De Jong, Ingram and Whiteman (1996) and provides the statistical foundations for the approach of Canova and De Nicolo (2003).

The computation of (7.21) is straightforward. $f(\mathsf{S}_\infty)$ can be obtained for each $\mathcal{M}_i, i = 1,2$. averaging S_∞ over draws of θ_i, given a draw of ϵ_t. $g(\mathsf{S}_{\infty,1}|y_{1t},\mathcal{M}_3)$ can be obtained with the techniques of chapter 9.

Exercise 7.30 *Continuing with the economy analyzed in exercise 7.28, consider two versions of the model: one where dividends follow a two-state Markov Chain and one where dividends follow a three-state Markov Chain.*
(i) Estimate a bivariate VAR for the US equity premium and the US real risk free rate. Produce 100 draws from a Normal-Wishart posterior distribution for the VAR coefficients (i.e. draw from a Wishart for Σ^{-1} *and, conditional on this draw, draw VAR coefficients from a normal with mean equal to the estimates and variance given by the draw of* Σ*).*
(ii) Assume $gy_1 = 0.9873, gy_2 = 1.0177, p_1 = 0.2, p_2 = 0.8, \rho_y = 0.8$ *and let* $\ln(\frac{\beta}{1-\beta}) \sim \mathsf{N}(3.476, 1.418^2), \varphi \sim \mathsf{N}(0.4055, 1.3077^2)$. *Also, assume that the growth rate of dividends in the third state is* $\ln(\frac{gy_3}{1-gy_3}) \sim \mathsf{N}(0.036, 1.185^2)$. *Draw 100 values for the parameters from these distributions and compute the equity premium and the risk free rate generated by the two models for each draw.*
(iii) Graphically examine the degree of overlap between the cloud of points generated by the two models and the cloud of points generated by the VAR.
(iv) Compute the posterior odds ratio (7.21) for the two models assuming that $\frac{g(\mathcal{M}_1|\mathcal{M}_3)}{g(\mathcal{M}_2|\mathcal{M}_3)} = 1$.
(v) Construct 68% contour probabilities from the posterior of $\mathsf{S} = (E(EP_t), E(R_t))$ *given the data. Provide a probabilistic assessment of the validity of the two models by counting the number of replications generating equity premium and the risk free rate within this contour.*

It is important to note that the procedure is conditional on $\mathcal{M}_3$, the empirical model bridging population moments and the data. Since VARs cam accurately represent economic data, they can be used to create this link. The procedure, however, is general: one could use more structural or more time series oriented specifications and could even employ "poor" models (as far as fit to the data is concerned), so long as the posterior of S_∞ is easy to compute.

7.5 Sensitivity of the measurement

Once the quality of a model is assessed and some confidence has been placed in its approximation to the data, the measurement or policy exercises one wants to perform can then be undertaken. In the simplest setup, the outcome of an experiment is a number (see e.g. Cooley and Hansen (1989)) and if one is interested in examining the sensitivity of the results to small variations in the neighbor of calibrated values, local sensitivity analysis can be undertaken informally, replicating the experiments for various parameter values or formally, calculating the elasticity of measurement with respect to variations in some of the components of θ (as in Pagan and Shannon (1985)).

When some uncertainty is allowed in the simulations, the outcome of the experiment is the realization of a random variable. Hence, one may be interested in assessing where the realization lies relative to the range of possible outcomes of the model. Some of the techniques outlined in section 7.4 can be used for this purpose. For example, one could construct simulated standard errors or confidence intervals, drawing vectors of parameters (and/or the stochastic processes for the exogenous variables) from some distribution (a-priori, empirical or sampling based). In this case the analysis is global in the sense that we analyze the sensitivity of the measurement to perturbations of the parameters over the entire range. Note also that in the approaches of Canova (1994), De Jong, Ingram and Whiteman (1996), (2000) and Geweke (1999), the evaluation procedure automatically and efficiently provides sensitivity analysis to global perturbations for the parameters within an economic reasonable range.

Besides simulation techniques, there are two alternative methods one can use to assess the sensitivity of the measurement. These approaches, initially suggested by Abdekhalem and Dufour (1998) for CGE economies, can be easily adapted to DSGE models. The first method is based on asymptotic expansions and formalizes Pagan and Shannon's (1985) local derivative approach.

Exercise 7.31 *Suppose* $\sqrt{T}(\theta_T - \theta) \to \mathsf{N}(0, \Sigma_\theta)$ *where* $det(\Sigma_\theta) \neq 0$.
(i) Show that if $h(\theta)$ *is* $m \times 1$ *vector of continuous and differentiable functions of* θ, $\sqrt{T}(h(\theta_T) - h(\theta)) \to \mathsf{N}(0, \Sigma_h = H(\theta)\Sigma_\theta H(\theta)')$ *where* $H(\theta) = \frac{\partial h(\theta)}{\partial \theta'}$.
(ii) Show that if rank$(H(\theta)) = m$, $T(h(\theta_T) - h(\theta))'\Sigma_h^{-1}(h(\theta_T) - h(\theta)) \xrightarrow{D} \chi^2(m)$. *Conclude that an asymptotic confidence set for* $h(\theta)$ *at the level of* $(1 - \varrho)$ *is* $CI_h(\theta) = \{h(\theta) : T(h(\theta_T) - h(\theta))'\Sigma_h^{-1}(h(\theta_T) - h(\theta)) \leq \chi^2_\varrho(m)\}$ *and* $P[h(\theta) \in CI_h(\theta)] = 1 - \varrho$.

Exercise 7.31 uses the asymptotic distribution of the parameters to construct confidence intervals for $h(\theta)$. Two drawbacks of this procedure are clear: first, we need to have an asymptotic distribution for the θ, which we typically do not have and, second, we need a model where the number of endogenous variables is equal to the dimension of θ. The second problem can be remedied by constructing rectangular (as opposed to ellipsoid) confidence sets for any $i = 1, \ldots, m$. That is, whenever $dim[h(\theta)] < m$ $CI_i(\theta_i) = \{h_i(\theta) : T\frac{(h_i(\theta_T) - h_i(\theta))^2}{\sigma_{ii}} \leq \chi^2(1)\}$ where $\sigma_{ii} = diag(\Sigma_{h_{ii}})$ and $P[h_i(\theta) \in CI_i(\theta_i)] = 1 - \varrho_i$. Note that a simultaneous confidence set not smaller than $1 - \varrho$ is obtained choosing ϱ_i so that $\sum_i \varrho_i = \varrho$ (e.g. $\varrho_i = \frac{\varrho}{m}$).

The second method does not employ asymptotic properties and only assumes that a set Θ with $P(\theta \in \Theta) \geq 1 - \varrho$ is available. This could be a prior or a posterior estimate if θ is random or a classical (small sample) confidence interval if Θ is random. Let $h(\Theta) = \{h(\theta_0) \in R^m : \text{ for at least some } \theta_0 \in \Theta\}$. Then $\theta \in \Theta$ implies that $h(\theta) \in h(\Theta)$ and $P[h(\theta) \in h(\Theta)] \geq P(\theta \in \Theta) = 1 - \varrho$. When h is nonlinear, direct computation of $P[h(\theta)]$ is difficult. As an alternative, let $h_i(\Theta) = \{h_i(\theta_0) \in R^m : \text{ for at least some } \theta_0 \in \Theta\}$. Then we can construct $P[h_i(\theta) \in h_i(\Theta), i = 1, \ldots, m] \geq 1 - \varrho$ and $P[h_i(\theta) \in h_i(\Theta)] \geq 1 - \varrho$, $i = 1, \ldots, m$. The first represents a simultaneous rectangular confidence set, the second a marginal rectangular confidence set. The following result establishes that these sets are intervals under general conditions.

Result 7.1 *If h is continuous and Θ compact and connected, each $h_i(\Theta)$ is compact and connected and $h_i(\Theta) = [h_i^{lo}(\Theta), h_i^{up}(\Theta)]$, $i = 1, 2, \ldots$ where $h_i^{lo} > -\infty, h_i^{up} < \infty$. (A set is connected if it is impossible to find two subsets O_1 and $O_2 \in R^m$ meeting O_3 such that $O_3 \subseteq O_1 \cup O_2$ and $O_3 \cap O_1 \cap O_2 = \emptyset$.)*

To find the upper and the lower limits of the interval one can use the following algorithm

Algorithm 7.5

1) *Construct $\Theta = \{\theta_0 \in R^m : (\theta - \theta_0)'\Sigma_\theta^{-1}(\theta - \theta_0) \leq \mathsf{C}(\theta)\}$ where $\Sigma_\theta = var(\theta)$ and C a function of θ.*

2) *Find the minimum and the maximum of $\mathsf{S}(\theta) = h_i(\theta_0) + \frac{\lambda}{2}[(\theta - \theta_0)'\Sigma_\theta(\theta - \theta_0) - \mathsf{C}(\theta)]$.*

3) *Set $\theta^{up} = argmax\ \mathsf{S}(\theta)$ and $\theta^{lo} = argmin\ \mathsf{S}(\theta)$.*

It is easy to verify that the first order conditions in 2) are $\frac{\partial h_i}{\partial \theta_0} - \lambda\Sigma_\theta(\theta - \theta_0) = 0$ and $(\theta - \theta_0)'\Sigma_\theta(\theta - \theta_0) - \mathsf{C}(\theta) = 0$. When Σ_θ is non singular, the θ^{up} and θ^{lo} that yield $h_i^{lo}(\Theta)$ and $h_i^{up}(\Theta)$ are $\theta_i = \theta \pm (\frac{\frac{\partial h_i}{\partial \theta_0}'\Sigma_\theta^{-05}\frac{\partial h_i}{\partial \theta_0}}{\mathsf{C}(\theta)})^{-1}\Sigma_\theta^{-1}\frac{\partial h_i}{\partial \theta_0}$. Note that the algorithm can be applied one dimension at the time, using rectangular intervals instead of an ellipsoid. Then $CI(\theta) = \{\theta \in R^m; \frac{(\theta - \theta_0)'\Sigma_\theta^{-1}(\theta - \theta_0)}{m} \leq F_\varrho\}$ is a confidence set for θ which contains 95% of the values. Also, we can knock out values which are incoherent with theory or do not give solutions to the model since $P(\theta \in \Theta) = P(\theta \in \Theta \cap \Theta_0) \geq 1 - \theta$, where Θ_0 is the set of

admissible values of θ. Finally, derivatives of the function h_i can be computed numerically, i.e. $\frac{\partial h(\theta)}{\partial \theta} = \frac{h(\theta+\iota)-h(\theta-\iota)}{2\iota}$, $\iota > 0$ and small.

Example 7.19 *Consider the economy described in example 2.4 of chapter 2, where all goods are cash goods and suppose we want to calculate the welfare costs of inflation. Cooley and Hansen (1989) showed that, depending on the average growth rate of money $\bar{M}$, the compensating variation in consumption needed to bring back consumers to the optimum varies between 0.107 to 7.59 percentage of GDP if the cash-in-advance binds for one quarter. Suppose that $\bar{M}$ is a random variable with mean 1.04 and standard deviation 0.01 (approximately the growth rate of money for the US over the 1970-2000 period). If money growth is normally distributed then, approximately, $h(\theta) \sim \mathsf{N}(0.21, 0.025)$. Hence a 68% confidence interval for the percentage of consumption in terms of steady state output needed to bring consumers back to their optimum is (0.185, 0.235).*

Exercise 7.32 *(Gourinchas and Jeanne) Consider a number of small open RBC economies. Population is growing at the rate $Pop_t = gp_{it} Pop_{t-1}$, where gp_{it} is country specific. The utility for country i is $\sum_\tau \beta^\tau Pop_{t+\tau} \frac{c_{t+\tau}^{1-\varphi}}{1-\varphi}$. Assume that $y_{it} = \zeta_{it} K_{it}^{1-\eta}$, where ζ_{it} is an AR(1) with persistence common across countries (equal to ρ_ζ) and that capital depreciates in each country at the rate δ. Assume that $\lim_{t\to\infty} gp_{it} = gp$ independent of i. Consider two situations: financial autharky and complete financial integration. In the former no international borrowing or lending occurs; in the latter countries can borrow at the rate $R_t = \frac{c_t^\varphi}{c_{t-1}^\varphi \beta}$, where $\frac{c_t}{c_{t-1}}$ is the gross growth rate of consumption under financial integration. Evaluate the gains of financial integration assuming $\beta = 0.96, \varphi = 2.0, \delta = 0.10, 1-\eta = 0.3$ and the steady state gross growth rate of consumption equals 1.012. Repeat the calculation assuming $1-\eta \sim \mathsf{U}[0.3, 0.5]$ (Hint: if $x \sim \mathsf{U}(a_1, a_2)$, $E(x) = 0.5(a_1+a_2)$, $var(x) = \frac{(a_2-a_1)^2}{12}$).*

7.6 Savings, Investments and Tax cuts: an example

Suppose we are interested in evaluating the effects of cuts in the income tax rate on investments and consumption and suppose we choose to study the issue with a two country RBC economy with complete markets. Baxter and Crucini (1993) claim that such a model can account for several features of the data, including the high correlation of domestic savings and domestic investments in open economies, without imposing restrictions on capital flows but use informal methods to reach this conclusion. Therefore, before undertaking the measurement of interest, we evaluate the quality of the model's approximation to the data using the techniques presented in this chapter. We assume that there is a single consumption good and labor is immobile. For each country $i = 1, 2$ preferences are given by: $E_0 \sum_{t=0}^{\infty} \frac{\beta^t}{1-\varphi} [C_{it}^\vartheta (1-N_{it})^{(1-\vartheta)}]^{1-\varphi}$ where C_{it} is private consumption, $1-N_{it}$ is leisure, β is the discount factor, $1-\vartheta(1-\varphi)$ the coefficient of relative risk aversion and ϑ the share of consumption in utility. Goods are produced according to $GDP_{it} = \zeta_{it}(K_{it})^{1-\eta}(X_{it}N_{it})^\eta$ $i = 1, 2$ where K_t is the capital, η is the share of labor in GDP, and $X_{it} = gnX_{it-1}$ $\forall i$ where $gn \geq 1$

captures the deterministic labor-augmenting technological progress. We let:

$$\begin{bmatrix} \ln\zeta_{1t} \\ \ln\zeta_{2t} \end{bmatrix} = \begin{bmatrix} \bar{\zeta}_1 \\ \bar{\zeta}_2 \end{bmatrix} + \begin{bmatrix} \rho_1 & \rho_2 \\ \rho_2 & \rho_1 \end{bmatrix} \begin{bmatrix} \ln\zeta_{1t-1} \\ \ln\zeta_{2t-1} \end{bmatrix} + \begin{bmatrix} \epsilon_{1t} \\ \epsilon_{2t} \end{bmatrix}$$

where $\epsilon_t = [\epsilon_{1t}\ \epsilon_{2t}]' \sim \mathsf{N}(0, \begin{bmatrix} \sigma_\epsilon^2 & \sigma_{12} \\ \sigma_{12} & \sigma_\epsilon^2 \end{bmatrix})$ and $[\bar{\zeta}_1, \bar{\zeta}_2]'$ is a vector of constants. Here σ_{12} controls the contemporaneous and ρ_2 the lagged spillover of the shocks. Capital goods are accumulated according to $K_{it+1} = (1-\delta_i)K_{it} + \frac{b}{2}(\frac{K_{it+1}}{K_{it}} - 1)^2 K_{it}$ $i = 1, 2$ where b is a parameter. Government expenditure is deterministic and financed with income taxes T_i^y and lump sum transfers T_{it}, $G_i = T_{it} + T_i^y GDP_{it}$. Finally, the resource constraint is:

$$\Psi(GDP_{1t} - G_{1t} - C_{1t} - K_{1t+1} + K_{it}) + (1-\Psi)(GDP_{2t} - G_{2t} - C_{2t} - K_{2t+1} + K_{it}) \geq 0 \quad (7.22)$$

where Ψ is the fraction of world population living in country 1. We first scale all variables by the labor augmenting technological progress, e.g. $gdp_{it} = \frac{GDP_{it}}{X_{it}}, c_{it} = \frac{C_{it}}{X_{it}}$, etc. and solve the model log-linearizing the optimality conditions around the steady state. The weights in the social planner problem are proportional to the number of individuals in each of the countries. Actual saving are computed as $Sa_t = GDP_t - C_t - G_t$. Data refers to the period 1970:1-1993:3 for US and for Europe; it is in real terms, seasonally adjusted and from OECD Main Economic Indicators. The properties of actual saving and investment are computed eliminating from the raw time series a linear time trend. The parameters of the model are $\theta = [\beta,\ \varphi,\ \vartheta,\ gn,\ \delta,\ \rho_1,\ \rho_2,\ \sigma_\epsilon,\ \sigma_{12},\ \Psi,\ b,\ T^y]$ plus steady state hours.

Parameter	Basic	Empirical Density	Subjective Density
Share of consumption ϑ	0.5	Uniform [0.3,0.7]	Normal (0.5, 0.02)
Steady State hours (N^{ss})	0.20	Uniform[0.2, 0.35]	Normal (0.2, 0.02)
Discount Factor (β)	0.9875	Truncated Normal [0.9855, 1.002]	Normal(0.9875, 0.01)
Utility Power (φ)	2.00	Truncated $\chi^2(2)[0, 10]$	Normal(2, 1)
Share of Labor in Output (η)	0.58	Uniform[0.50, 0.75]	Normal(0.58, 0.05)
Growth rate (gn)	1.004	Normal(1.004, 0.001)	1.004
Depreciation Rate of Capital (δ)	0.025	Uniform[0.02, 0.03]	Normal(0.025, 0.01)
Persistence of Disturbances (ρ_1)	0.93	Normal(0.93, 0.02)	Normal(0.93, 0.025)
Lagged Spillover (ρ_2)	0.05	Normal(0.05, 0.03)	Normal(0.05, 0.02)
Standard Deviation of Technology Innovations (σ_z)	0.00852	Truncated $\chi^2(1)$ [0, 0.0202]	Normal(0.00852, 0.004)
Contemporaneous Spillover (σ_{12})	0.40	Normal(0.35, 0.03)	Normal(0.4, 0.02)
Country Size (Ψ)	0.50	Uniform[0.10, 0.50]	0.5
Adjustment cost to capital (b)	1.0	1.0	1.0
Tax Rate (T^y)	0.0	0.0	0.0

Table 7.4: Parameters selection

The exogenous processes are the two productivity disturbances so that $\epsilon_t = [\ln\zeta_{1t},\ \ln\zeta_{2t}]'$. We generate samples of 95 observations to match the actual data and the number of replications is 500. We evaluate the quality of the model using the diagonal elements of the 4×4

spectral density matrix of the data (savings and investment for the two countries) and the coherence between saving and investment in the two countries. Spectral density estimates are computed smoothing periodogram ordinates with a flat window. In the benchmark parametrization the θ vector is the same as in Baxter and Crucini (1993) (see the first column of table 7.4) except for σ_ϵ which we take from Backus, Kehoe and Kydland (1995), and ϑ, which does not appear in their specification. We also allow for parameter uncertainty using the approaches of Canova (1994) and of De Jong, Ingram and Whiteman (1996). In the first case empirical based distributions are constructed using existing estimates or, when there are none, choosing a-priori an interval and assuming a uniform distribution. In the second case distributions are normal, with means equal to the calibrated parameters and dispersions a-priori chosen. The distributions are displayed in the second and third columns of table 7.4. A comparison of the model and the data at business cycle frequencies (3-8 years) is in table 7.5. The first two rows report the average spectral densities and coherences at business cycle frequencies for actual and simulated data when parameters are fixed. The next two rows report Watson's average measure of fit at business cycle frequencies. The first is obtained minimizing the variance of saving and investment in country 1 and the second, minimizing the variance of savings and investment in country 2.

	US Spectra		Europe Spectra		US Coherence	Europe Coherence
	Sa	Inv	Sa	Inv	Sa-Inv	Sa-Inv
Actual data	0.75	0.88	0.68	0.49	85.41	93.14
Simulated data	0.36	0.18	0.35	0.18	94.04	93.00
Watson						
Identification 1	0.02	0.05	0.20	0.23	0.04	0.13
Identification 2	0.24	0.21	0.05	0.04	0.20	0.15
Covering						
Fixed parameters	46.46	8.63	55.71	43.57	98.99	92.91
Subjective density	35.30	23.40	32.89	37.00	98.17	90.34
Empirical density	19.63	18.60	21.11	20.20	94.71	95.69
Critical Value						
Fixed parameters	90.80	99.89	82.16	93.91	15.60	49.04
Subjective density	71.80	89.90	66.00	76.60	19.80	51.89
Empirical density	62.50	79.70	73.30	74.60	33.46	29.60
Error						
Fixed parameters	0.25	0.55	0.30	0.28	-9.17	0.37
Subjective density	0.19	0.56	0.29	0.28	-9.01	0.81
Normal density	0.13	0.58	0.42	0.35	-6.07	-2.86

Table 7.5: The Fit of the Model

National saving is highly correlated with domestic investment in both areas and the average coherence at business cycle frequencies is higher for Europe than for the US. The variability of both US series is higher and US investment is almost two times more volatile than the European one. Because the model is symmetric, the variability of simulated saving and investment is similar in the two countries, but low relative to the data. However,

consistent with the data the variability of national savings is higher than that for domestic investment. Consistent with Baxter and Crucini's claims, the model produces high national saving and investment correlations at business cycle frequencies. In fact, the model coherences for the US are higher than those found in the actual data. Watson's measures suggest that, on average, the size of the error at business cycle frequencies is between 2% and 5% of the spectral density of those variables whose variance is minimized and between 20% and 25% of the spectral density of other variables. Changes in the coherences across identifications are somewhat relevant and the model fits them better when we minimize the variance of US variables.

The next three rows (Covering) report how many times on average, at business cycle frequencies, the diagonal elements of the spectral density matrix and the coherences of model generated data lie within a 95% confidence band for the corresponding statistics of actual data. Clearly, a number close to 95% indicates a "good" model performance. We compute 95% confidence bands for the actual data in two ways: using asymptotic theory and using a version of the parametric bootstrap procedure of Diebold, Ohanian and Berkowitz (1998). In this latter case, we run a four variable VAR with 6 lags and a constant, construct replications for saving and investment for the two countries, bootstrapping the residuals of the VAR model, estimate the spectral density matrix of the data for each replication and extract 95% confidence bands after ordering the replications, frequency by frequency. Replications for the time series generated by the model are constructed using Monte Carlo techniques in three different ways: keeping the parameters fixed at the values displayed in the first column of table 7.4 or randomizing them using draws from the distributions listed in the second and third columns of table 7.4. Since results are similar we only report probability coverings using an asymptotic 95% band. This third set of statistics confirms that the model matches coherences better than volatilities at business cycle frequencies and that the covering properties of the model do not improve when parameters uncertainty is allowed.

Under the heading "Critical Value" we report the percentile of the simulated distribution of the spectral density matrix of saving and investment in the two countries where the value of the spectral density matrix of actual data (taken here to be estimated without an error) lies, on average, at business cycle frequencies. Values close to 0% (100%) show poor fit - the actual spectral density matrix is in the tail of the distribution of the spectral density matrix of simulated data - while values close to 50% should be considered good. Also here we report a case with fixed parameters and two with random ones.

With fixed parameters the model generates average coherences which are much higher than in US data but close to the median for Europe (actual values are in the 15th and 50th percentile). With random parameters (and empirical based priors), the situation improves for the US (actual coherence moves up to the 33rd percentile) but not for Europe. Also, with fixed parameters the model generates a distribution for variability which is skewed to the left and only partially overlaps a normal asymptotic range of variabilities for the data. Parameter uncertainty, by tilting and stretching the shape of the simulated distribution, ameliorates the situation.

Finally we computed the distributional properties of the approximation error, i.e. we compute the distribution of the error needed to match the spectral density matrix of the actual data, given the model's simulated spectral density matrix. To do this we draw at each replication, parameters and innovations from the posterior distribution of the VAR representation of the actual data, construct time series of interest and estimate the spectral density matrix of the four series. At each replication, we also draw parameters and innovations from the distributions presented in table 7.4, construct the spectral density matrix of simulated data and compute $\mathcal{S}_v^l(\omega) = \mathcal{S}_y^l(\omega) - \mathcal{S}_x^l(\omega)$ at each $l = 1, \dots L$. If the model replicates the DGP, the distribution for this error would be degenerate at each frequency. Otherwise, features of this distribution (median value, skewness, kurthosis, etc.) may help to pin point what is missing from the model. The last three rows in table 7.5 ("Error") present the median (across replications) of the average error at business cycle frequencies for the six statistics. The first row reports results when parameters are fixed and the next two when parameters are randomized. The results are similar in the three cases: the model fails to generate enough variability at business cycle frequencies for US investments while for the other three variables the error is smaller. The results for coherences depend on the country. For the US, the model generates systematically higher coherences (negative spectral errors) while for Europe the opposite is true.

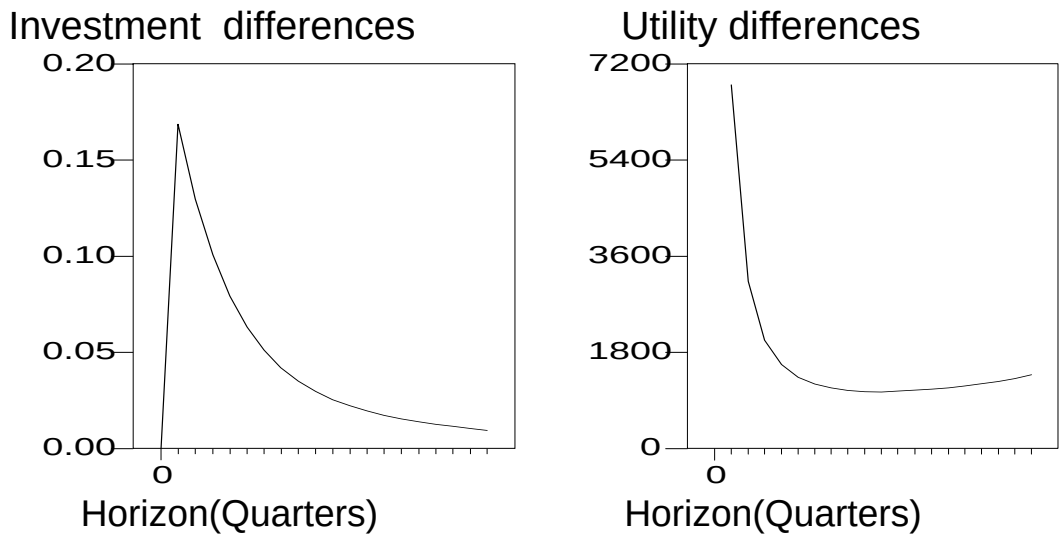


Figure 7.4: Effects of tax cuts

In conclusion, in agreement with Baxter and Crucini (1993), the model generates high coherence between national saving and investment at business cycle frequencies. Its magnitude is similar to the one observed in European data, but uniformly higher than the one observed in US data, regardless of whether parameters are fixed or random. However,

the model has hard time to account for the variability of saving and investment in both countries at business cycle frequencies.

To measure the effects of tax cuts we perform two simulations: one with a tax rate of 0.20 and one with a tax rate of 0.0, using the parameters listed in the first column of table 7.4 and ask how large is the difference in investment responses along the adjustment path when a positive productivity shock hits the domestic economy. Figure 7.4 reports graphically this difference in percentage terms: investment response is significantly larger without taxes in the first few periods but the gains dissipate reasonably fast. The utility differences induced by these two paths are also significant, but level off after about 5 periods. In fact, the compensating variation in consumption needed to restore the utility level obtained with no taxes is 0.11 each period, about 14% of steady state consumption. The magnitude of this number is robust. For example, the lower bound to the level of compensating variations obtained for $\vartheta \in [0.3, 0.7]$ and $\varphi \in [1, 4]$ is 0.09 each period.

Chapter 8: Dynamic Macro Panels

Panels of macroeconomic time series are now used in many fields. For example, in studying the transmission of shocks, one would like to have a cross country point of view. Similarly, when examining convergence of income per capita of nations or regions, one would like to account for both cross sectional and time series interactions.

The models we consider in this chapter borrow from the micro panel literature in the sense that the specifications do not allow for lagged interdependencies across units. This is an important shortcoming: since interdependencies are the results of world market integration they can hardly be neglected in applied macro work. In chapter 10 we study how to introduce them in dynamic panel models using a Bayesian point of view. The setup we use here is different from the standard treatments of panel data since models are explicitly dynamic, either because of the presence of lagged dependent variables or of lags of exogenous variables. For a comprehensive account of existing approaches when models are static see Hsiao (1989), Baltagi (1995) or Hayashi (2001).

The econometric theory developed in the context of micro panels is somewhat inappropriate for macro applications. Estimators are typically constructed for samples which have a small time series (T) and large cross section (n). Therefore the properties of estimators are derived exploiting asymptotics in the cross section. In macro panels, typically, neither n nor T are large and, most of the times, $T > n$. This should be kept in mind when deciding the estimator to use and the inference one is allowed to make. Another crucial problem in macro data is dynamic heterogeneity. In micro panels, even when the model is dynamic, no slope heterogeneity is allowed for and unit specific characteristics are mostly captured with a time invariant fixed or random effect. In macro panels this restriction is, in general, inappropriate: heterogeneous dynamic reflects policies or regulations and one wants to be able to evaluate differences, if they emerge. Note also that for most of this chapter, we consider models which are stationary or display time invariant structures. We alter this setup in chapter 10 where panel VAR models with time varying coefficients are examined.

We start in the next section with an example to motivate our interest in dynamic panel analysis. In section 2 we consider panel (VAR) models with no slope heterogeneities but with unit specific factors; we describe how to estimate them using instrumental variables; illustrate the problems that traditional fixed and random effect estimators encounter in this specification; examine how to construct estimates of the unit specific (time invariant) effect and, finally, how to test interesting hypotheses. In section 3 we introduce slope

heterogeneities, describe a series of estimators for this type of models, study their properties and propose a test to detect slope heterogeneities. In section 4 we describe approaches to pooling and examine their pros and cons. In many situations single unit (time series) estimates can be improved upon pooling the information coming from the cross section, even when no interdependencies are allowed for. Finally, in the last section we use some of the methods presented in this chapter to examine whether money is superneutral in the long run long in the cross section of G-7 countries.

8.1 From economic theory to dynamic panel data

To motivate our use of dynamic panels in macroeconomic analysis we consider the problem of modelling growth in open economies. Barro, Mankiw and Sala (1995) presented an extension of the standard Solow model which has interesting insights from a theoretical point of view and important empirical implications.

We have a set of countries, indexed by i, which are small, in the sense that they take the world interest rate as given, and accumulate two types of capital, human and physical. The representative agent in each country i maximizes $\sum_t \beta^t \frac{c_{it}^{1-\varphi}}{1-\varphi}$ subject to the constraint

$$c_{it} + K_{it+1} + hk_{it+1} + sa_{it+1} \leq \zeta_t^{\eta_i} K_{it}^{\eta_k} hk_{it}^{\eta_{hk}} + (1-\delta_k)K_{it} + (1-\delta_{hk})hk_{it} + (1+r_t)sa_{it} \quad (8.1)$$

where K_{it} is physical capital, hk_{it} is human capital and sa_{it} are lending to (borrowing from) the rest of the world; ζ_t represents total factor productivity and its efficiency, measured by η_i, may differ across i. We assume that each country i has limited borrowing capacity. In particular, $-sa_{it} \leq K_{it}$ while hk_{it} cannot be used as collateral in international borrowing. When the constraint is binding, capital and borrowing are perfect substitutes in the portfolio of agents and $1 + r_t = (1 - \delta_k) + \eta_k \frac{GDP_{it}}{K_{it}}$, which implies that

$$K_{it} = [(1 + r_t) - (1 - \delta_k)]^{-1} \eta_k GDP_{it} \quad (8.2)$$

Using (8.2) in the production function we have $GDP_{it} = \zeta_{it}^{\dagger} hk_{it}^{\eta_1}$ where $\eta_1 = \frac{\eta_{hk}}{1-\eta_k}$ and $\zeta_{it}^{\dagger} = [\frac{\zeta_t^{\eta_i} \eta_k^{\eta_k}}{[(1+r_t)-(1-\delta_k)]^{\eta_k}}]^{1/1-\eta_k}$. Maximizing utility with respect to (c_{it}, hk_{it}) and using (8.2) in the resource constraint yields the following two equilibrium conditions

$$\begin{aligned} hk_{it+1} &= (1-\eta_k)\zeta_{it}^{\dagger} hk_{it}^{\eta_1} + (1-\delta_{hk})hk_{it} - c_{it} & (8.3) \\ c_{it}^{-\varphi} &= \beta E_t[c_{it+1}^{-\varphi}(\eta_{hk}\zeta_{it+1}^{\dagger} hk_{it+1}^{\eta_1 - 1} + (1-\delta_{hk})] & (8.4) \end{aligned}$$

Exercise 8.1 *Verify that in the steady states* $c_i^{ss} = (1 - eta_k)\zeta_i^{\dagger}(hk^{ss})^{\eta} - \delta_{kh} hk_i^{ss}$ *and* $hk_i^{ss} = [\frac{1-\beta(1-\delta_{kh})}{\beta \eta_{hk} \zeta_i^{\dagger}}]^{(1-\eta_{hk})/(\eta_{hk}-1+\eta_k)}$ *and are different across* i *if* $\eta_i \neq \eta_{i'}, i \neq i'$.

ii) Verify that log linearizing (8.3)-(8.4) and setting $\psi_{i1} = (1-\eta_k)\eta_1 \zeta_i^{\dagger}(\frac{GDP_i}{hk_i})^{ss} + (1-\delta_{hk})$, $\psi_{i2} = \frac{(1-\eta_k)(\eta_1-1)\zeta_i^{\dagger}(hk_i^{ss})^{\eta_1-1}}{(1-\eta_k)\zeta_i^{\dagger}(hk_i^{ss})^{\eta_1-1} + (1-\delta_{hk})}$

$\psi_{i3} = \frac{(1-\eta_k)\zeta_i^{\dagger}(hk_i^{ss})^{\eta_1-1}}{(1-\eta_k)\zeta_i^{\dagger}(hk_i^{ss})^{\eta_1-1}+(1-\delta_{hk})}$ *we have (in percentage deviations from the steady states)*

$$\hat{hk}_{it+1} = \psi_{i1}\hat{hk}_{it} + (1-\eta_k)(\frac{GDP_i}{hk_i})^{ss}\hat{\zeta}_{it}^{\dagger} - \frac{c_i^{ss}}{hk_i^{ss}}\hat{c}_{it} \tag{8.5}$$

$$-\varphi\hat{c}_{it} = -\varphi E_t\hat{c}_{it+1} + \psi_{i2}E_t\hat{hk}_{it+1} + \psi_{i3}E_t\hat{\zeta}_{it+1}^{\dagger} \tag{8.6}$$

Letting $y_{it} = [c_{it}, hk_{it}]$ and adding an expectational error to equation (8.6) to capture differences between actual and expected values of $\hat{c}_{it+1}$, $\hat{\zeta}_{it+1}^{\dagger}$, $\hat{hk}_{it+1}$ we can rewrite (8.5)-(8.6) as a vector of first order difference equations for each i of the form $\mathcal{A}_{i0}\hat{y}_{it+1} = \mathcal{A}_{i1}\hat{y}_{it} + \mathcal{A}_{i2}\hat{e}_{it}$ where $\hat{e}_{it}$ is a function of $\hat{\zeta}_{it}^{\dagger}$ and of the expectational error $\hat{v}_{it}$. Letting $\bar{y}_i$, $\bar{e}_i$ be the steady state values of y_i and e_i and adding them back we have

$$\mathcal{A}_{i0}y_{it+1} = \mathcal{A}_{i1}y_{it} + \varrho_i + \epsilon_{it} \tag{8.7}$$

where $\varrho_i = (\mathcal{A}_{i0} - \mathcal{A}_{i1} - \mathcal{A}_{i2})\bar{y}_i$, $\epsilon_{it} = \frac{\mathcal{A}_{i2}\bar{y}_i}{\bar{e}_i}e_i$.

Equation (8.7) is a bivariate VAR(1) model for each i, with unit specific fixed effects and heterogeneous dynamics. The model of this section implies, in general, that whenever the steady states are different across units, the dynamics leading to the steady state will also be different. Therefore, models of this type deliver the framework examined in section 3. There are two special cases of equation (8.7) which can be of interest. The first obtains when dynamics are homogenous and there are unit specific fixed effects. In the model we have used, it is clear that this is possible if and only if β is different across units (it is the only parameter which appears in the steady state but not in the dynamics) and if $\eta_i = \eta_i'$, i.e. Total factor Productivity (TFP) has the same efficiency across units. Such a model will be dealt with in section 2. A second special case of interest emerges when fixed effects are absent and the dynamics are heterogeneous. This is possible when $E_t\zeta_{it+1}^{\dagger}$ differs across i (e.g. because expectations are different). Finally, it is worth noting that, by construction, there is no interaction across units. This is entirely due to the small open economy assumption. For example, if a world budget constraint is added to the problem important interactions across units would emerge. Hence, the panel VAR models with interdependencies considered in chapter 10 can be originated, e.g., from a two country model with an international budget constraint for borrowing and lending.

Exercise 8.2 *Consider a basic RBC model and suppose that government expenditure provides utility to the agents and that private and public consumption are substitutes in the utility function. Assume that the instantaneous utility function for country i is* $u(c_{it}, G_{it}, N_{it}) = (c_{it} + \vartheta_g G_{it})^{\vartheta}(1-N_{it})^{1-\vartheta}$, *that the budget constraint is* $c_{it} + K_{it+1} + G_{it} = \zeta_t^{\eta_i}K_{it}^{1-\eta}N_{it}^{\eta} + (1-\delta_k)K_{it}$, *that* $G_{it} = G_t + a_{ig}\zeta_t$ *and that expenditure is financed by lump sum taxation on a period-by-period basis, where* G_t *is an iid stochastic process and* a_{ig} *is a parameter which regulates the response of country i expenditure to the state of the home technology.*
i) Write down the Euler equation for the problem for each i and log linearize it.
ii) Under what conditions would the vector of log-linearized Euler equations produce a panel with homogeneous dynamics and country specific intercept or a panel with heterogeneous dynamics and no fixed effects?

8.2 Panels with Homogeneous dynamics

The model we consider in this section has the form

$$y_{it} = A_{0t} + \sum_{j=1}^{\infty} A_{1jt} y_{it-j} + \sum_{j=1}^{\infty} A_{2jt} x_{it-j} + A_{3t}\varrho_i + e_{it} \tag{8.8}$$

where e_{it} is a martingale difference process with covariance matrix Σ_i, y_{it} is a $m_1 \times 1$ vector for each $i = 1, \ldots, n$, $t = 1, \ldots, T$, x_{it} is a $m_2 \times 1$ vector of exogenous variables, ϱ_i is the (unobservable) unit specific effect and for each j, A_{1jt} is a $m_1 \times m_1$ matrix and A_{2jt} a $m_1 \times m_2$ matrix.

Equation (8.8) is general: lagged dependent and exogenous variables appear on the right hand side and, in principle, time varying coefficients are allowed for. Furthermore, heterogeneities are possible both in the level and in the variance. One important restriction, which will be relaxed later on, is that the dynamics are identical across units. This restriction allows us to construct estimators of the parameters using cross sectional information at each t and permits the use of standard asymptotic theory when testing hypotheses, even when y_{it} is non-stationary. We also assume that x_{it} includes, or may be composed entirely of, variables which are common across units. Notice that we have modelled ϱ_i as a fixed effect. While in micro panels one has the choice between fixed and random effects, in macro data a fixed effect specification is preferable for two reasons. First, if ϱ_i captures omitted variables, it is likely to be correlated with the regressors (a possibility typically excluded by a random effect specification). Second, a macro panel in general contains all the units of interest and thus is less likely to be a random sample from a larger population (e.g. an OECD panel typically includes all the OECD countries). Note that since e_t is a martingale difference $E(x_{it-\tau}e_{it}) = E(y_{it-\tau}e_{it}) = 0$ for all $\tau < 0$ and $E(\varrho_i e_{it}) = 0$, for all i.

Equation (8.8) is not estimable since ϱ_i is unobservable. In a static model, one eliminates this fixed effect by subtracting averages from (8.8) and estimating the model in deviations from the means with OLS. In the next exercise we ask the reader to verify that estimates obtained from the transformed model are consistent in a setup where ϱ_i is unobservable and may be correlated with other regressors.

Exercise 8.3 *Consider the model $y_{it} = x_{it}A_2 + \varrho_i + e_{it}$ where $i = 1, \ldots, n$, $t = 1, \ldots, T$, $E[e_{it}|x_{it}] = 0$, $E[e_{it}^2|x_t] = \sigma_e^2$, $E[e_{it}, e_{i'\tau}] = 0\ \forall i \neq i'$, $\tau \neq t$; $E[\varrho_i|x_{it}] \neq 0$ and $E[e_{it}|\varrho_i] = 0$.*
(i) Show that OLS estimates of the parameters are inconsistent.
(ii) Show that consistent estimates can be obtained taking deviations from individual means, i.e running the OLS regression $y_{it} - \bar{y}_i = (x_{it} - \bar{x}_i)A_2 + (e_{it} - \bar{e}_i)$, $\bar{y}_i = \frac{1}{T}\sum_t y_{it}$; $\bar{x}_i = \frac{1}{T}\sum_t x_{it}$; $\bar{e}_i = \frac{1}{T}\sum_t e_{it}$. Show that coefficients which are constant for each i, in every t cannot be estimated.
(iii) Assume that $E[e_{i,t}|x_{i,t}] \neq 0$. Write down a 2SLS estimator for the specifications in (i) and (ii) assuming that $E[e_{i,t}|z_{i,t}] = 0$ where $z_{i,t}$ is a set of instruments. Show that a 2SLS estimator is consistent in the original model but not necessarily so in the transformed one (Hint: $E[\bar{e}_i|z_{it}] \neq 0$ even if $E[e_{it}|z_{i,t}] = 0$).

Exercise 8.3 shows a peculiar result: OLS in the transformed model is consistent but 2SLS, in general, is not. To insure consistency of a 2SLS estimator we need to strengthen the orthogonality condition and impose that $E((e_{i,\tau} - \bar{e}_i)|z_{i,t}) = 0 \ \forall t, \tau$.

The case of ρ_i correlated with regressors is common in macroeconomics. For example, suppose one is interested in studying the effects of money on inflation across countries. Clearly, the path of money supply may be related to other country specific characteristics ϱ_i, (for example, the stance of fiscal policy) therefore making potential regressors correlated with the individual effect.

Example 8.1 *(Growth and volatility) Theoretically, it is unclear what the sign of the relationship between growth and volatility should be: volatility could be a manifestation of the adoption of new technologies that induce cost restructuring and this could provide a positive link between the two. The relationship could also be negative as volatility can result in wasted human capital or deter investment. Letting the growth rate of value added be ΔGDP_{it}, the volatility of the growth rate of value added be x_{1it} and the growth rate of other regressors be x_{2t}, then a typical model studied in the empirical literature is $\Delta GDP_{it} = A_0 + x_{1it}A_1 + x_{2t}A_2 + \varrho_i + e_{it}$. Since ϱ_i are unobservable, they are typically pooled together with e_{it} into an error term. Note that OLS cannot be used to estimate A_1 and A_2 if, e.g., ϱ_i captures political factors, since in units were instability is strong, volatility may be high and growth low so the residuals are negatively correlated with the regressors. Alternatively, if i refers to sectors and ϱ_i captures industry specific technological breakthroughs, the residuals will be positively correlated with the regressors. Imbs (2002) presents estimates of the parameters obtained using deviations from time means and the UNIDO data base, when x_{2t} is either omitted or when it is not, it measures competitiveness. The data refers to 15 OECD countries, covers the sample 1970-1992 and has a maximum of 28 sectors for each country. Estimates are obtained when i represents a sector-country combination.*

Specification	A_1	A_2	A_0	R^2
1	*4.893*		*0.121*	*0.02*
	(2.63)		*(3.68)*	
2	*5.007*	*-0.059*	*0.133*	*0.02*
	(2.66)	*(-0.39)*	*(2.94)*	

Table 8.1: Growth and Volatility

The relationship between volatility and growth is statistically positive and economically significant. For example, in the first regression a one percent increase in volatility increases the average yearly sectorial output growth by 0.5 percent. Note also that comparative advantage for the sector-country pairs is insignificant once fixed effects are taken into account. Finally, the explanatory power of both regressions is small: volatility has only a marginal explanatory power for value added growth.

Exercise 8.4 *Consider the model $y_{it} = \bar{y} + \varrho_i + \varrho_t + \alpha x_{it} + e_{it}, i = 1, \ldots, n$ where ϱ_t is a time effect and suppose $\sum_i \varrho_i = 0$, $\sum_t \varrho_t = 0$. Suppose you estimate this model using a dummy*

variable for each i and a time trend. Show that OLS estimates of (ϱ_i, α) are consistent if T is large. Show that estimates of ϱ_i are inconsistent for large n. Show that, for large n, is it better to assume that ϱ_i is a random variables with mean ϱ and variance σ^2_ϱ.

8.2.1 Pitfalls of standard methods

When lagged dependent variables are present and the time series dimension of the panel is small or fixed, taking deviations from the mean does not produce consistent estimates.

Example 8.2 *We illustrate the problems existing in this case using a version of equation (8.8) where $m_1 = 1$, $A_{0t} = x_{it} = 0, \forall t$, $A_{1jt} = A_{1j}$ and $A_{1j} = 0, \quad j \geq 2$ and $A_{3t} = 1, \forall t$. Hence (8.8) reduces to an AR(1) model with unit specific fixed effects. We assume y_{io} fixed and var$(e_{it}) = \sigma^2$. A pooled estimator for A_1 is $A_{1p} = \frac{\sum_{i=1}^{n}\sum_{t=1}^{T}(y_{it}-\bar{y}_i)(y_{it-1}-\bar{y}_{i,-1})}{\sum_{i=1}^{n}\sum_{t=1}^{T}(y_{it-1}-\bar{y}_{i,-1})^2} = A_1 + \frac{\sum_{i=1}^{n}\sum_{t=1}^{T}(e_{it}-\bar{e}_i)(y_{it-1}-\bar{y}_{i,-1})/nT}{\sum_{i=1}^{n}\sum_{t=1}^{T}(y_{it-1}-\bar{y}_{i,-1})^2/nT}$ where $\bar{y}_{i,-1}$ is the mean of y_{it-1}. Repeatedly substituting into the model and summing over t we have $\sum_t y_{it-1} = \frac{1-A_1^T}{1-A_1}y_{io} + \frac{(T-1)-TA_1+A_1^T}{(1-A_1)^2}\varrho_i + \sum_{j=0}^{T-2}\frac{1-A_1^{T-1-j}}{1-A_1}e_{i,1+j}$. Since $E(\varrho_i e_{it}) = 0$*

$$plim_{n\to\infty}\frac{1}{nT}\sum_i\sum_t(e_{it}-\bar{e}_i)(y_{it-1}-\bar{y}_{i,-1}) = -plim_{n\to\infty}\frac{1}{n}\sum_i\bar{y}_{i,-1}\bar{e}_i = -\frac{\sigma_e^2}{T^2}\frac{(T-1)-TA_1+A_1^T}{(1-A_1)^2} \quad (8.9)$$

$$plim_{n\to\infty}\sum_i\sum_t(y_{it-1}-\bar{y}_{i,-1})^2 = \frac{\sigma_e^2}{1-A_1^2}(1-\frac{1}{T}-\frac{2A_1}{(1-A_1)^2}\frac{(T-1)-TA_1+A_1^T}{T^2}) \quad (8.10)$$

For consistency we need that (8.9) converges to zero and that (8.10) converges to a fixed number. As $T \to \infty$, (8.9) does indeed go to zero and (8.10) goes to $\frac{\sigma_e^2}{1-A_1^2}$. However if T is fixed, the estimator is inconsistent, even when $n \to \infty$.

Exercise 8.5 *Show that the asymptotic bias of A_{1p} in the model of example 8.2 is $plim_{n\to\infty}(A_{1p}-A_1) = -\frac{1+A_1}{T-1}(1-\frac{1}{T}\frac{1-A_1^T}{1-A_1})[1-\frac{2A_1}{(1-A_1)(T-1)}(1-\frac{1-A_1^T}{T(1-A_1)})]^{-1}$. Show that for T large, $p\lim_{n\to\infty}(A_{1p}-A_1) \approx \frac{-(1+A_1)}{T-1}$.*

Intuitively, the bias appears because to eliminate ϱ_i from the model we have introduced a correlation of order $1/T$ between the explanatory variable and the residual of the model $(y_{it} - \bar{y}) = A_1(y_{it-1} - \bar{y}_{i,-1}) + (e_{it} - \bar{e}_i)$. In fact, $\bar{y}_{i,-1}$ is correlated with $(e_{it} - \bar{e}_i)$ even if e_{it} are serially uncorrelated since $\bar{e}_i$ contains e_{it-1} which is correlated with y_{it-1}. When T is large the right hand side variables are uncorrelated with the errors, but for T small, estimates of the mean effects are biased and this bias is transmitted to the estimates of A_1.

Table 8.2 shows that if $A_1 > 0$, the bias is generally negative and not negligible. For highly persistent processes, like those typically observed in macro time series, the bias is about 15 percent when T=20 and still 6 percent when T=40.

	$A_1 = 0.2$	$A_1 = 0.5$	$A_1 = 0.8$	$A_1 = 0.95$
T=10	-0.1226	-0.1622	-0.2181	-0.2574
T=20	-0.0607	-0.0785	-0.1044	-0.1300
T=30	-0.0403	-0.0516	-0.0672	-0.0853
T=40	-0.0302	-0.0384	-0.0492	-0.0629

Table 8.2: Bias in the AR(1) coefficient

Example 8.3 *(Production function estimation) One typical case where problems with lagged dependent variables exist is in estimating production functions across sectors. Let* $GDP_{it} = N_{it}^{\eta_N} K_{it}^{\eta_k} \zeta_{it}$ *where, in principle,* $\eta_N + \eta_k \neq 1$ *and where the technological progress* ζ_{it} *is parametrized as* $\ln \zeta_{it} = \bar{\zeta}_i + A_1 \ln \zeta_{it-1} + e_{it}$*. Taking logs of the production function and quasi-differencing, we have* $\ln GDP_{it} = A_1 \ln GDP_{it-1} + \eta_N(\ln N_{it} - A_1 \ln N_{it-1}) + \eta_k(\ln K_{it} - A_1 \ln K_{it-1}) + \bar{\zeta}_i + e_{it}$*. Hence, unless* ζ_t *is iid, estimation of* η_k, η_N *using production functions in deviation from the mean will produce biased estimates, even when* n *is large.*

The problem described in example 8.2 is generic and it is present even when cross sectional techniques (as opposed to pooled techniques) are used to estimate the parameters.

Exercise 8.6 *(Nickell) Consider the cross sectional OLS estimator* A_{1t} *obtained using only the t-th cross section* $A_{1t} = \frac{\sum_{i=1}^{n}(y_{it-1} - \bar{y}_{i,-1})(y_{it} - \bar{y}_i)}{\sum_{i=1}^{n}(y_{it-1} - \bar{y}_{i,-1})^2}$ *where* $\bar{y}_{i,-1}$ *is the mean of* y_{it-1}*.*
(i) Show that $p\lim_{n\to\infty}(A_{1t} - A_1) = -\frac{1+A_1}{T-1}[1 - A_1^{t-1} - A_1^{T-t} + \frac{(1-A_1^T)}{T(1-A_1)}][1 - \frac{2A_1}{(T-1)(1-A_1)}(1 - A_1^{t-1} - A_1^{T-t} + \frac{(1-A_1^T)}{T(1-A_1)})]^{-1}$ *(this is the same as the bias obtained in exercise 8.5).*
(ii) Argue that the inconsistency of A_{1t} *is of order* $(1/T)$*; that its bias depends on which cross section* t *is used and that it is smaller at the end of the sample.*

The standard alternative to demeaning the variables is to use a random effect estimator. Although we have argued that such an approach is conceptually problematic for macro data, we show that treating ϱ_i as random does not solve the inconsistency problem in models with lagged dependent variables.

Example 8.4 *Suppose we move* ϱ_i *into the error term and construct a pooled estimator* $\tilde{A}_{1p} = A_1 + \frac{\sum_{i=1}^{n}\sum_{t=1}^{T}(e_{it} + \varrho_i)y_{it-1}/nT}{\sum_{i=1}^{n}\sum_{t=1}^{T}(y_{it-1})^2/nT}$*. The numerator of this expression can be written as* $\frac{1}{T}\frac{1-A_1^T}{1-A_1}cov(y_{io}, \varrho_i) + \frac{1}{T}\frac{\sigma_\varrho^2}{(1-A_1)^2}((T-1) - TA_1 + A_1^T)$ *and the denominator is* $\frac{1-A_1^{2T}}{T(1-A_1)^2}\sum\frac{y_{i0}^2}{n} + \frac{\sigma_\varrho^2}{(1-A_1)^2}\frac{1}{T}(T - 2\frac{1-A_1^T}{1-A_1} + \frac{1-A_1^{2T}}{1-A_1^2}) + \frac{2}{T(1-A_1)}(\frac{1-A_1^T}{1-A_1} - \frac{1-A_1^{2T}}{1-A_1^2})cov(\varrho_i, y_{i0}) + \frac{1}{T}\frac{\sigma_e^2}{(1-A_1^2)^2}[(T-1) - TA_1^2 - A_1^{2T}]$*. If* y_{io} *is fixed the covariance term drops out of the expression (otherwise, it would be positive (any guess why?)), but the numerator is different from zero even when* $T \to \infty$ *and is larger, the larger is the variance of the unit specific effects* σ_ϱ^2*.*

Exercise 8.7 *Consider the model* $y_{it} = A_1 y_{it-1} + A_2 x_{it} + \varrho_i + e_{it}$ *and let* $\tilde{y}_{it} = y_{it} - \bar{y}_i$, $\tilde{y}_{it-1} = y_{it-1} - \bar{y}_{i,-1}$, $\tilde{x}_{it} = x_{it} - \bar{x}_i$, $\tilde{e}_{it} = e_{it} - \bar{e}_i$.
(i) Show that, using a pooled OLS estimator on the demeaned model, we obtain

$$\begin{aligned} A_{1p} &= A_1 + (\tilde{y}'_{-1}(I - \tilde{x}(\tilde{x}'\tilde{x})^{-1}\tilde{x}')\tilde{y}_{-1})^{-1}\tilde{y}'_{-1}(I - \tilde{x}(\tilde{x}'\tilde{x})^{-1}\tilde{x}')\tilde{e} \\ A_{2p} &= A_2 - (\tilde{x}'\tilde{x})^{-1}\tilde{x}'\tilde{y}_{-1}(A_{1p} - A_1) + (\tilde{x}'\tilde{x})^{-1}\tilde{x}'\tilde{e} \end{aligned}$$

(ii) Show that

$$\begin{aligned} p\lim_{n\to\infty}(A_{1p} - A_1) &= (p\lim_{n\to\infty}\frac{1}{nT}\tilde{y}'_{-1}(I - \tilde{x}(\tilde{x}'\tilde{x})^{-1}\tilde{x}')\tilde{y}_{-1})^{-1}(p\lim_{n\to\infty}\frac{1}{nT}\tilde{y}'_{-1}\tilde{e}) \\ p\lim_{n\to\infty}(A_{2p} - A_2) &= -[p\lim_{n\to\infty}(\tilde{x}'\tilde{x})^{-1}\tilde{x}'\tilde{y}_{-1}]p\lim_{n\to\infty}(A_{1p} - A_1) \end{aligned}$$

Exercise 8.7 shows that the bias in A_{2p} depends on the bias in A_{1p} and the relationship between the exogenous variables $\tilde{x}$ and the lagged endogenous variables $\tilde{y}_{-1}$, both in deviations from their mean. If $E(xy_{-1}) > 0$ the bias in A_{2p} is positive (recall that the bias in A_{1p} is negative).

It is important to stress that disregarding dynamic effects does not help. In fact, if the true model has lagged dynamics and a static model is estimated, the error term will be correlated with the regressors and this correlation will remain even after variables are demeaned.

Exercise 8.8 *Suppose* $y_{it} = \rho_i + A_1 y_{it-1} + A_2 x_t + e_{it}$ *and one estimates* $y_{it} = \rho_i + A_2 x_{it} + v_{it}$ *where* $v_{it} = e_{it} + A_1 y_{it-1}$. *Show that* v_{it} *is correlated with the regressor if* x_{it} *is serially correlated. Show that demeaning the estimated model does not eliminate this correlation.*

One implication of exercise 8.8 is clear: running a static demeaned regression and correcting for serial correlation is unlikely to produce consistent estimates of the parameters when the model for each unit is dynamic - the standard case with macroeconomic time series.

8.2.2 The Correct approach

To deal with unobservable variables when lagged dependent variables are present define $\xi_t = \frac{A_{3t}}{A_{3t-1}}$ and quasi-difference (8.8) to get,

$$y_{it} = A^+_{0t} + \sum_{j=1}^{q_1+1} A^+_{1jt} y_{it-j} + \sum_{j=1}^{q_2+1} A^+_{2jt} x_{t-j} + e^+_{it} \tag{8.11}$$

where $A^+_{0t} = A_{0t} - \xi_t A_{0t-1}$, $A^+_{11t} = \kappa_1 + A_{11t}$; $A^+_{1jt} = A_{1jt} - \xi_t A_{1,j-1,t-1}$; $A^+_{1q_1+1t} = -\xi_t A_{q_1,t-1}$, $A^+_{21t} = A_{21t}$; $A^+_{2jt} = A_{2jt} - \xi_t A_{2j-1,t-1}$; $A^+_{2q_2+1t} = -\xi_t A_{2q_2 t-1}$, $e^+_{it} = e_{it} - \xi_t e_{it-1}$. If $A_{3t} = A_3, \forall t$, (8.11) is simply the differenced version of (8.8) and the approach to eliminate

the unobserved fixed effect corresponds to the one suggested by Anderson and Hsiao (1982). Note that in (8.11) the orthogonality conditions are $E(x_{t-\tau}e_{it}^{+}) = E(y_{it-\tau}e_{it}^{+}) = 0$ for all $i,\ \tau > 0$. The Anderson and Hsiao estimator, which was designed for the AR(1) version of this model with no x_{it}, uses y_{it-2} or $(y_{it-2} - y_{it-3})$ as instruments to estimate A_{j1}^{+}. Note that, because of differencing, y_{it-1} is not a valid instrument - it is correlated with the error term.

Exercise 8.9 *Suppose* $q_1 = 1, q_2 = 0, A_{0t} = A_0, A_{1t} = A_1, A_{2t} = 0, \forall t, A_{3t} = A_3 = 1$. *Display a IV estimator for the parameters and describe the instruments you would use. Give conditions that insure consistency when* $n \to \infty$, *when* $T \to \infty$ *or both.*

Since the orthogonality conditions are valid for any $\tau > 0$, there are many instruments one could use; the Anderson and Hsiao estimator uses one particular set of instruments but, as we have seen in chapter 5, we can improve by appropriately combining all available information. In the case of constant coefficients the derivation of a GMM-style estimator is a straightforward application of the ideas described in chapter 5.

Example 8.5 *(Arellano and Bond) Let* $\Delta y_i = [\Delta y_{i2}, \ldots, \Delta y_{it}]$, $z_{i\tau} = [1, y_{i\tau-2}, \ldots, y_{i1}, x_{\tau-2}, \ldots, x_1]$ $z_i = \text{diag}[z_{i1}, \ldots, z_{it}]$ $Z = [z_1', \ldots, z_n']'$; $X_i = [\Delta y_{it-2}, \ldots, \Delta y_{i1}, \Delta x_{t-1}, \ldots, \Delta x_1]$, $X = \text{diag}\ [X_1, \ldots, X_n]$, $\Delta y = \text{diag}\ [\Delta y_1, \ldots, \Delta y_n]$. *Then a GMM-style estimator for* $\alpha = [A_{11}, \ldots A_{1q_1}, A_{21}, \ldots, A_{2q_2}]'$ *is* $\alpha_{GMM} = (X'ZWZ'X)^{-1}(X'ZWZ'\Delta y)$ *where* W *is a weighting matrix. As in chapter 5, the optimal* W *depends on the covariance of the instruments. An estimator for* W *is* $\hat{W} = (\frac{1}{n}\sum_i Z_i\Omega Z_i)^{-1}$ *where* Ω *is a* $(T-2)\times(T-2)$ *matrix with 2 on the main diagonal, -1 on the first subdiagonals and zero elsewhere.*

When the coefficients are time varying, little more work is needed. We start by giving the conditions for identification in the original and the transformed model.

Exercise 8.10 *Show that the order condition for identification of the parameters of the transformed model is* $T > (max(q_1, q_2)) + 3$. *Show that the order condition for identification of the parameters of the original model is* $T > 3 * max(q_1, q_2) + 2$.

The next exercise adapts the results of exercise 8.10 to two important special cases.

Exercise 8.11 *Show that if* $\xi_t = 1$, *the order condition for identification of the original parameters is* $T > 2 * max(q_1, q_2) + 2$. *Show that if the original parameters are time invariant, the order condition for the identification is* $T > max(q_1, q_2) + 2$.

Intuitively, more data points are needed to be able to pin down the parameters when the model is nonstationary.

Example 8.6 *Suppose* $y_{it} = A_1 y_{it-1} + \varrho_i + e_{it} - \phi e_{it-1}$ *and suppose* $T = 4$. *Then the model in first difference for each* t *is* $\Delta y_{i4} = A_1\Delta y_{i3} + e_4 - \phi e_3$; $\Delta y_{i3} = A_1\Delta y_{i2} + e_3 - \phi e_2$; $\Delta y_{i2} = A_1\Delta y_{i1} + e_2 - \phi e_1$. *Since* $q = 1$, $T \geq q+3 = 4$ *and since there are* $\frac{(T-q-2)(T-q-1)}{2} = 1$ *restrictions, to estimate the AR coefficient we have to use* y_{i1} *as an instrument for* y_{i3} *when* $T = 4$ *is considered. The equations for* $T = 3$ *and* $T = 2$ *are not estimable.*

The time series of a macro panel are typically of uneven quality due to differences in recorded practices or statistical procedures. Therefore, it is important to understand what happens when (y_{it}, x_t) are measured with error. Suppose for this purpose that $x_t^c = x_t + \epsilon_t^x$, $y_{it}^c = y_{it} + \epsilon_{it}^y$ where $E(e_{it}\epsilon_{it}^y) = E(e_{it}\epsilon_t^x) = 0$ and that the measurement errors are iid and uncorrelated with the true value of the series.

Exercise 8.12 *Consider the version of (8.8) where $A_{1jt} = A_{1j}, A_{2jt} = A_{2j}, A_{0t} = 0,\ \forall t$ but where both y_{it} and x_t are measured with error. Show that the system has the form $\Delta y_{it} = \sum_j A_{1j}\Delta y_{it-1} + \sum_j A_{2j}\Delta x_t + v_{it}$ where Δ is the differencing operator and $v_{it} = \Delta e_{it} + \Delta\epsilon_{it}^y + \sum_j A_{1j}\Delta\epsilon_{it-j}^y + \sum_j A_{2j}\Delta\epsilon_{t-j}^x$. Let $z_{it} = [1, y_{it-q-2}^c, \ldots, y_{i1}^c, x_{it-q-2}^c, \ldots, x_{i1}^c]$ where $q = max\{q_1, q_2\}$. Show that z_{it} is uncorrelated with v_{it}. Show that the order condition for identification is $T \geq 2q + 2$.*

Since the presence of (classical) measurement error introduces a MA structure in the error term, efficiency can be improved if this structure is taken into account in the estimation process. Consistency is not affected.

To estimate time varying parameters let $q = max\{q_1, q_2\}$, $y_t = [y_{it}, \ldots y_{nt}]', x_t = [y_{it}, \ldots y_{nt}]', E_t = [e_{it}, \ldots, e_{nt}]', X_t = [1, y_{t-1}, \ldots, y_{t-q_1-1}, x_{t-1}, \ldots, x_{t-q_2-1}]$, $\alpha_t = [A_{0t}^+, A_{11t}^+, \ldots, A_{2,q_1+1,t}^+, A_{21t}^+, \ldots, A_{2,q_2+1,t}^+]'$. Then (8.8) can be written in simultaneous equation format as $y_t = X_t\alpha_t + E_t$ and stacking the $T - q - 2$ observations we have

$$y = X\alpha + E \tag{8.12}$$

Let $z_t = [1, y_{t-2}, \ldots, y_1, x_{t-2}, \ldots, x_1]$. Clearly z_t changes with t. Let $z = \text{diag}[z_{q+3}, \ldots, z_t]$. For the instruments to be valid we need $plim_{n\to\infty}\frac{Z'E}{n} = 0$ (this is a $(T-q-2)n \times 1$ vector of conditions). Then using the logic of GMM, α can be estimated with a two-step approach.

Exercise 8.13 *Describe a two step approach to estimate α. Show that a 2SLS estimator in this case is given by $\alpha_{2SLS} = [X'Z'W^{-1}Z'X]^{-1}X'Z'W^{-1}Z'y$ where $W_{\tau t} = \sum_{i=1}^n e_{it}e_{i\tau}Z_{it}'Z_{i\tau}$, and e_{it} is the (i,t) element of E. Is α_{2SLS} efficient?*

As usual, consistent estimates of e_{it} can be used in the formula for $W_{\tau t}$, e.g. $e_{it,(2sls)} = y - X\alpha_{2sls}$.

It is worthwhile to examine in detail GMM estimation when there are no exogenous variables and the dynamics are restricted to be AR(1) since several empirical applications (convergence exercises, production function estimation, growth accounting) fit into this framework, if the left hand side variables are appropriately scaled.

Example 8.7 *Consider the model $y_{it} = A_1 y_{it-1} + \varrho_i + e_{it}$ where $|A_1| < 1$ and $E(e_{it}) = E(e_{it}e_{i\tau}) = 0\ \forall t \neq \tau$, T fixed and n large. Suppose we wish to estimate A_1 absent any distributional information on ϱ_i and e_{it}. Given the assumptions made, y_{it-2} is a valid instrument for the estimation of A_1 in the model in first differences. For $T \geq 3$, there are $(T-2)(T-1)/2$ linear moment restrictions of the type $E[(e_{it} - e_{it-1})y_{it-\tau}] = 0$ where $t = 3, \ldots T; \tau = 2, \ldots t-1$. For example, if T=4, there are three orthogonality restrictions*

$E[(e_{i4}-e_{i3})y_{i2}]=0, E[(e_{i4}-e_{i3})y_{i1}]=0$ *and* $E[(e_{i3}-e_{i2})y_{i1}]=0$. *Rewrite the restriction as* $E[z_i'\Delta e_{it}]=0$ *where* z_i *is a* $(T-2)\times(T-2)(T-1)/2$ *block diagonal matrix whose* τ*-th block is* $(y_{it},\dots,y_{i\tau})$ *(i.e.* $z_i=\text{diag}\{y_{i1},\dots y_{i\tau}\}, \tau=1,\dots T-2$*).*
A GMM estimator of A_1 *is based on the sample counterpart of* $E(z_i'Deltae_{it}]$*, i.e.* $\frac{1}{n}\sum_{i=1}^{n} z_i'(e_i-e_{i,-1})=n^{-1}Z'\Delta e$ *where* $\Delta e=e-e_{-1}=((e-e_{-1})_1,\dots,(e-e_{-1})_n)'$ *is a* $n(T-2)\times 1$ *vector and* $Z=(Z_1,\dots,Z_n)$ *is a* $n(T-2)\times(T-2)(T-1)/2$ *matrix. Then:*

$$A_{1,GMM}=argmin_{A_1}(\Delta E'Z)W_n(Z'\Delta E)=\frac{\Delta y_{-1}'ZW_nZ'\Delta y}{\Delta y_{-1}'ZW_nZ'\Delta y_{-1}} \tag{8.13}$$

where y_{-1} *indicates lagged variables,* $\Delta y=y-y_{-1}$ *and* W_n *is a weighting matrix.*

Exercise 8.14 *(i) Using the appropriate central limit theorem argue that* $\Sigma_n^{-0.5}n^{-0.5}Z'\Delta E\rightarrow \mathcal{N}(0,1)$ *where* Σ_n *is the average(over the cross section) covariance matrix of* $z_i'\Delta e_i$*.*
(ii) Show that with the assumptions made, Σ_n *can be replaced by* $\hat{\Sigma}_n=n^{-1}\sum_{i=1}^{n}(z_i'\hat{\Delta e}_i\hat{\Delta e}_i z_i)$ *where* $\hat{\Delta e}_i=\Delta y_i-\hat{A}_1\Delta y_{i,-1}$ *and* $\hat{A}_1$ *is a preliminary consistent estimate.*
(ii) Show that a consistent estimate of the asymptotic covariance matrix of $A_{1,GMM}$ *is* $\hat{var}(A_{1,GMM})=n\frac{\Delta y_{-1}'ZW_n\hat{\Sigma}_nW_n\hat{\Sigma}_nZ'\Delta y_{-1}}{\Delta y_{-1}ZW_nZ'\Delta y_{-1}}$.

As in chapter 5 we can derive the optimal choice for W_n by minimizing $\hat{var}(A_{1,GMM})$.

Exercise 8.15 *(i) Show that a one-step estimator is obtained setting* $W_n=(n^{-1}\sum_i z_i'\Omega z_i)^{-1}$ *where* Ω *is a* $(T-2)\times(T-2)$ *matrix with* $+2$ *on the main diagonal,* -1 *on the first sub-diagonals and zero otherwise.*
(ii) Show that $W_n=\hat{\Sigma}_n^{-1}$ *is optimal (it produces an estimator we denote by* $A_{1,2step}$*).*
(ii) Show that $A_{1,GMM}$ *and* $A_{1,2step}$ *are asymptotically equivalent if* e_{it} *are independent and homoschedastic across* n *and* T*.*

Clearly, since IV estimation is inefficient relative to GMM estimation, the Anderson-Hsiao estimator of A_1 obtained by regressing Δy_{it} on Δy_{it-1} using either Δy_{it-2} or y_{it-2} as instruments is inefficient relative to the GMM estimators derived in example 8.7 and exercises 8.14 and 8.15.

8.2.3 Restricted models

At times in the estimation process, one would like to consider (linear) restrictions of the form $\alpha_t=R\theta_t+r$, where $dim(\theta_t)<\dim(\alpha_t)$. Restrictions of this type may be theory based or may simply come from stationarity constraints. It is relatively easy to estimate restricted models and test the validity of these restrictions. Before describing the machinery necessary to do this, we present an example where such restrictions may occur.

Example 8.8 *Consider the case of a group of small open economies which take world interest rate as given. Suppose we are interested in examining the effect of capital tax*

rebates on investments using a model like (8.8) where the world interest rate is included in x_{it}. Suppose that some of these economies are dollarized and some of them are not. In this case, it may make sense to model interdependencies within each group but not among groups. Hence, there are restrictions on the A_{1jt} matrices one should take into account in the estimation process.

Following the steps we have used in chapter 4, define $Y_t^\dagger \equiv Y_t - X_t r = X_t R\theta_t + E_t \equiv X_t^\dagger \theta_t + E_t$ and assume that $E(Z'E) = 0$ and $E(Z'X^\dagger) \neq 0$. Then a GMM estimator for $\theta = (\theta_1, \ldots, \theta_t)$ is $\theta_{GMM} = [(X^\dagger)'ZW^{-1}Z'X^\dagger]^{-1}((X^\dagger)'ZW^{-1}Z'Y^\dagger)$.

To test the validity of the restrictions one could use any of the tests described in chapter 5. For example, let $n \times \mathsf{S}_{un,t} = (Y_t - X_t\alpha_{t,GMM})Z_t W_n^{-1} Z_t'(Y_t - X_t\alpha_{t,GMM})$ and $n \times \mathsf{S}_{re,t} = (Y_t^\dagger - X_t^\dagger\theta_{t,GMM})Z_t W_n^{-1} Z_t'(Y_t^\dagger - X_t^\dagger\theta_{t,GMM})$. Let $\alpha = (\alpha_1, \ldots, \alpha_t)$ and $\theta = (\theta_1, \ldots, \theta_t)$. Using standard asymptotic arguments, it follows that for $n \to \infty$, $\mathsf{S}_{un,t} \xrightarrow{D} \chi^2(dim(Z) - dim(\alpha))$ and $\mathsf{S}_{re,t} \xrightarrow{D} \chi^2(dim(\alpha) - dim(\theta))$. Hence, as $n \to \infty$, the likelihood ratio statistics $LR_t = \mathsf{S}_{re,t} - \mathsf{S}_{un,t} \xrightarrow{D} \chi^2(dim(Z) - dim(\theta))$, $t = 1, \ldots, T$.

Exercise 8.16 *Assume $\alpha_t = \alpha \ \forall t$. Describe how to implement a Wald test for the hypothesis $\alpha = R\theta + r$.*

As in VAR models one may be interested in testing a series of hypotheses and proceed at each stage conditional on the results obtained at the previous stage. For example, one would like to test how many lags should be included in the model and, conditional on the results, test some economic restriction, such as long run neutrality or steady state convergence. As in chapter 4, the significance level needs to be appropriately adjusted to take into account the sequential testing approach.

Example 8.9 *Let the first restriction be $\alpha_t = R\theta_t + r$ and the second $\theta_t = \bar{R}\phi_t + \bar{r}$. Let $n \times \mathsf{S}_{un,t} = (Y_t - X_t\alpha_{t,GMM})Z_t W_n^{-1} Z_t'(Y_t - X_t\alpha_{t,GMM})$, $n \times \mathsf{S}_{re_1,t} = (Y_t^\dagger - X_t^\dagger\theta_{t,GMM})Z_t W_n^{-1} Z_t'(Y_t^\dagger - X_t^\dagger\theta_{t,GMM})$ $n \times \mathsf{S}_{re_2,t} = (Y_t^\ddagger - X_t^\ddagger\phi_{t,GMM})Z_t W_n^{-1} Z_t'(Y_t^\ddagger - X_t^\ddagger\phi_{t,GMM})$ where $Y_t^\ddagger \equiv Y_t^\dagger - X_t\bar{r} = X_t\bar{R}_1\phi_t + E_t \equiv X_t^\ddagger\phi_t + E_t$ Define $LR_{1,t} = \mathsf{S}_{re_1,t} - \mathsf{S}_{un,t}$; $LR_{2,t} = \mathsf{S}_{re_2,t} - \mathsf{S}_{re_1,t}$, where the latter is a test of the second set of restrictions, conditional on the first set being true. If a_j is the significance of test $j = 1, 2$, a test of the second hypothesis has significance $a_1 + a_2 - a_1a_2$. Hence, for $a_1 = a_2 = 0.10$, the significance level of the second restrictions, conditional on the first being correct, is 0.19.*

These testing ideas can be used to examine whether there is heterogeneity in levels among units. From a practical point of view this is important since if $\varrho_i = \varrho \ \forall i$ and the parameters are time invariant, across sectional/pooled OLS estimates of the parameters of interest are consistent. However, if $\varrho_i \neq \varrho_{i'}$, first differencing and instrumental variables are needed. This distinction allows us to design a GMM-type test for the hypothesis of interest.

Example 8.10 *Consider a univariate model $y_{it} = \varrho_i + A_1 y_{it-1} + e_{it} = A_1 y_{it-1} + \epsilon_{it}$. A_{1p} obtained pooling the cross sections is inconsistent since ϵ_{it} is correlated with $y_{it-\tau}$ for all τ.*

First differencing the specification we have $\Delta y_{it} = A_1 \Delta y_{it-1} + \Delta \epsilon_{it}$. *Since* $E(y_{it-\tau}, \Delta\epsilon_{it}) = 0$, $\tau \geq 2$, y_{it-2} *is a valid instrument. Suppose* $T = 3$. *If* $\varrho_i = \varrho \forall i$, *there are three orthogonality conditions* $E(y_{i2}\epsilon_{i3}) = E(y_{i1}\epsilon_{i3}) = E(y_{i1}\epsilon_{i2}) = 0$, $\forall i$ *which can be used to estimate one (common) AR parameter. The last two conditions imply* $= E(y_{i1}(\epsilon_{i3} - \epsilon_{i2})) = 0$. *Since this condition holds both under the null and the alternative it can be employed to estimate* A_1. *The other two conditions,* $E(y_{i2}\epsilon_{i3}) = E(y_{i1}\epsilon_{i2}) = 0$, $\forall i$ *are valid only under the null. Therefore, given an estimate of* A_1, *they can be used to test whether an individual effect is present, given an estimate of* A_1.

A general formulation of the testing idea contained in example 8.10 is the following. Let $y_{it} = \sum_{j=1}^{q_1} A_{1j} y_{it-j} + \varrho_i + e_{it} = \sum_{j=1}^{q_2} A_{1j} y_{it-j} + \epsilon_{it}$. Then under the null $E(y_{it-j}\epsilon_{it}) = 0$ for $j = 1, \ldots, T$, $t = q_1 + 1, \ldots, T$. Under the alternative $E(y_{it-j}\Delta\epsilon_{it}) = 0$ for $j = 1, \ldots, T$, $t = q_1 + 2, \ldots T$, but $E(y_{it-j}\epsilon_{it}) \neq 0$. Given q_1 lags and T observations, there are $0.5 * (T(T-1) - q_1(q_1-1))$ orthogonality conditions. Since there are q_1 parameters to be estimated under the null, there are $\nu = 0.5 * (T(T-1) - q_1(q_1-1)) - q_1$ overidentifying restrictions. Therefore $S = \frac{(Y - \sum_j A_{1j} Y_{-j}) Z W^{-1} Z' (Y - \sum_j A_{1j} Y_{-j})}{n} \rightarrow \chi^2(\nu)$.

Exercise 8.17 *Suppose* $y_{it} = \sum_{j=1}^{q_1} A_{1j} y_{it-j} + \sum_{j=1}^{q_2} A_{2j} x_{t-j} + \varrho_i + e_{it}$ *where* $E(x_{t-\tau} e_{it}) = 0$, *for* $\tau = 1, \ldots T, t = q+1, \ldots T$ *where* $q = max(q_1, q_2)$. *How many orthogonality conditions are there? How many degrees of freedom has the test for homogeneity in this case?*

In time invariant models, it is typical to use only a subset of the orthogonality conditions, since the information contained in e.g. $E(z_{t-\tau}, e_{it})\tau$ large, may be negligible. In this case let j be the number of covariances of interest and let $jT - 0.5 * j(j+1) - 0.5 * q_1(q_1+1)$ be the number of orthogonality conditions. If $j > q_1$, the orthogonality conditions in an AR(q_1) model under the null are:

$$E(y_{it-\tau}\Delta e_{it}) = 0 \quad \tau = 2, \ldots, t-1, \quad t = (q_1+2), \ldots, j \tag{8.14}$$

$$E(y_{it-\tau}\Delta e_{it}) = 0 \quad \tau = 2, \ldots, j, \quad t = (j+1), \ldots T \tag{8.15}$$

$$E(y_{iq_1+1-\tau} e_{iq_1+1}) = 0 \quad \tau = 1, \ldots, q_1 \tag{8.16}$$

$$E(y_{it-1} e_{it}) = 0 \quad t = (q_1+2), \ldots, T \tag{8.17}$$

Here (8.14)-(8.15) hold under the null and the alternative; (8.16)-(8.17) hold only under the null. As usual, employing a limited number of instruments produces a less efficient test.

Exercise 8.18 *Show the conditions that need to be satisfied if* $j \leq q_1$.

Exercise 8.19 *Consider the model* $y_{it} = A_1 y_{it-1} + \varrho_i + e_{it}$; *let* $T = 4$ *and* $j = 2$.
i) Write down the orthogonality conditions implied by the model, distinguishing between those valid under both hypotheses and those valid only under the null.
ii) Stack the equations for all time periods and write $Y = A_1 Y_{-1} + e$. *Using* $Z = \text{diag}(z_1, \ldots z_n)$, *construct an IV variable estimator for* A_1.
iii) Derive a GMM estimator for A_1.
iv) Write down a J-style test for the overidentifying restrictions.

It is important, to stress that a GMM-style test for heterogeneity is inappropriate if some time series have a unit root. In that case, one should use likelihood ratio tests, which have good properties even when unit roots are present (see e.g. Smith and Fuertes (2003, p. 30)).

8.2.4 Recovering the individual effect

In macro applications it is important to obtain estimates of ϱ_i's and have a feeling of their cross sectional distribution, since these parameters may capture differences in national policies and/or other cross unit characteristics. When first differences are taken, ϱ_i is non-identifiable from the estimated specification. Nevertheless, it is easy to obtain an estimate of it. Let $\hat{\alpha}$ be an estimator of α obtained from the model in first differences. Let $\hat{\epsilon}_{it} = y_{it} - x_{it}\hat{\alpha}$. Taking time series averages $\hat{\epsilon}_i = \frac{1}{T}\sum_{t=1}^{T}\hat{\epsilon}_{it} = \bar{y}_i - \hat{\alpha}\bar{x}_i$ where $\bar{y}_i = \frac{1}{T}\sum_t y_{it}$, $\bar{x}_i = \frac{1}{T}\sum_t x_{it}$. Since $\frac{1}{T}\sum_{t=1}^{T}\hat{e}_{it} \to E(e_i) = 0$ as $T \to \infty$, $\hat{\epsilon}_i = \hat{\varrho}_i = \bar{y}_i - \hat{\alpha}\bar{x}_i$.

Example 8.11 *We have estimated a panel AR(1) model with country specific intercepts using quarterly real GDP data for 11 European nations for the sample 1988:1-2003:3. We have taken first differences to estimate the common AR parameter, pooling the data, using 11 lags as instruments and averaged over T the residuals for each i. Figure 8.1 shows the distribution of country specific effects which is clearly skewed, somewhat leptokurtic. We have tested for homogeneity of the individual effects (assuming they are all equal to the mean), with the test described in the previous subsection. The test has a p-value of 0.07, indicating that heterogeneities are somewhat important. However, if we exclude Austria and Finland, homogeneity is not rejected.*

8.2.5 Some Practical issues

There are at least three issues of practical interest worth discussing when estimating models with homogeneous dynamics and unit specific fixed effects. First, we have seen that OLS estimates of the (common) AR parameters are biased when the model is dynamic. Hence, how large should T be before the bias becomes negligible? Second, we know that GMM is more efficient than IV based on a single instrument, but also that estimates of the weighting matrix converge very slowly, therefore making the outcome unpredictable. What can we say about the trade off between bias and efficiency in GMM estimators? Finally, what is the trade-offs between OLS and IV? In particular, how large are the relative biases of the two estimators for panels of the typical size found in macroeconomics? To answer these questions we have constructed experimental data using

$$\begin{aligned} y_{it} &= A_1 y_{it-1} + A_2 x_{it} + \varrho_i + e^y_{it} \\ x_{it} &= A_3 x_{it-1} + e^x_{it} \end{aligned} \tag{8.18}$$

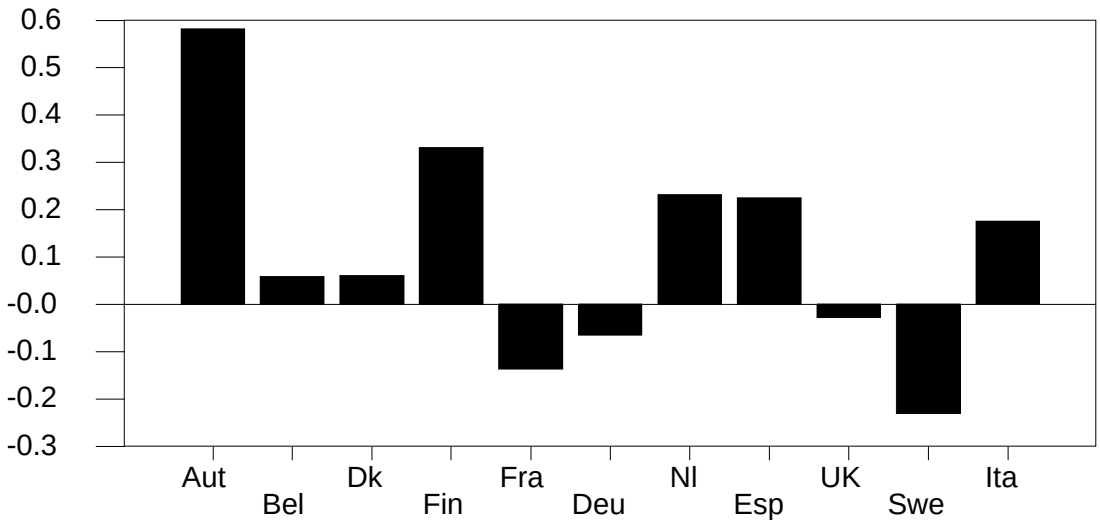


Figure 8.1: Individual Effects, GDP

where $e_{it}^y \sim \ iid\ \mathsf{N}(0,\sigma_y^2)$, $e_{it}^x \sim \ iid\ \mathsf{N}(0,\sigma_x^2)$. We set $A_2 = 1 - A_1$ so that changes in A_1 affect the short run dynamics but not the long run relationship between x and y. The parameters controlling the experiments are $A_1, A_3, \sigma_y^2, \sigma_x^2$. In the first experiment, we set $\sigma_x = 1, A_3 = 0.5, \sigma_y = 2$ and we let A_1 vary with T. We also set $y_{io} = x_{i0} = 0$ and discard the first 100 observations. We perform 500 replications for each combination of parameters using n=100. Table 8.3 reports the results. In parenthesis, are numerical standard errors.

T	A_1	Bias in A_1	Bias in A_2
10	0.2	-0.059 (0.025)	0.017 (0.026)
	0.8	-0.232 (0.033)	0.004 (0.043)
20	0.2	-0.027 (0.015)	0.010 (0.018)
	0.8	-0.104 (0.018)	0.006 (0.026)
40	0.2	-0.017 (0.011)	0.007 (0.014)
	0.8	-0.056 (0.013)	0.006 (0.023)

Table 8.3: Monte Carlo evidence I

The bias in A_1 is typically more severe than the bias in A_2: it increases with the value of A_1 and decreases with T. Note that the bias is about 10 percent when $A_1 = 0.8$ and 20 data points are available but it increases to about 30 percent when T= 10. When time series are persistent, the bias is significant even when $T = 40$.

In the second experiment we let n vary with T and A_1. We focus on GMM estimators,

obtained using two and five instruments, using both one-step and two-step approaches.

Two important conclusions can be drawn from table 8.4. First, the bias induced by the estimation of the optimal weighting matrix is significant and the one-step estimator is always best. Notice that the bias in the two-steps estimator increases, surprisingly, with T and, as expected, is larger the larger is the AR coefficient. Second, using two instruments typically produces smaller biases. However, it is also the case that with five instruments, the bias is more precisely estimated.

			GMM one-step		GMM Two-stepa	
T	n	A_1	2 Instuments	5 Instruments	2 Instruments	5 Instruments
10	20	0.2	-0.041 (0.066)	-0.050 (0.056)	-0.043 (0.081)	-0.077 (0.102)
		0.8	-0.222 (0.124)	-0.241 (0.115)	-0.249 (0.168)	-0.336 (0.198)
10	100	0.2	-0.011 (0.035)	-0.012 (0.022)	-0.009 (0.036)	-0.011 (0.032)
		0.8	-0.056 (0.071)	-0.079 (0.059)	-0.056 (0.072)	-0.081 (0.066)
20	20	0.2	-0.032 (0.044)	-0.038 (0.038)	-0.084 (0.118)	-0.263 (0.199)
		0.8	-0.137 (0.081)	-0.144 (0.066)	-0.441 (0.281)	-0.880 (0.498)
20	100	0.2	-0.005 (0.022)	-0.007 (0.019)	-0.005 (0.025)	-0.008 (0.027)
		0.8	-0.028 (0.040)	-0.039 (0.039)	-0.030 (0.039)	-0.048 (0.040)
40	20	0.2	-0.022 (0.034)	-0.026 (0.032)	-0.188 (0.148)	-0.423 (0.363)
		0.8	-0.108 (0.059)	-0.111 (0.044)	-0.837 (0.294)	-1.154 (0.509)
40	100	0.2	-0.003 (0.018)	-0.005 (0.013)	-0.004 (0.017)	-0.017 (0.028)
		0.8	-0.024 (0.030)	-0.030 (0.025)	-0.031 (0.036)	-0.089 (0.049)

Table 8.4: Monte Carlo evidence II

Comparing tables 8.3 and 8.4, one can see that GMM estimators perform better when n is large but, for a fixed n, their performance is far from appealing. Also, estimators become less biased when T increases, except when n is small. Overall, GMM and OLS biases are similar, when using a one-step estimator and $T = 40$.

8.3 Dynamic heterogeneity

So far we have examined panels where the dynamics are homogeneous across units. However, there are many situations when the homogeneity assumption is not particularly attractive and dynamic heterogeneities should be allowed for. For example, in growth theory it has become common to empirically study issues of convergence and polarization of income distributions (see Barro and Sala (1992), Quah (1996), Canova and Boldrin (2001)) and policy circles are often interested in predicting the (long run) effects of certain policy choices across units of a panel. Alternatively, it is often emphasized that political economy issues may shape the dynamics of government debt (see e.g. Alesina and Perotti (1995)). Finally, in many situations researchers care if market forces or policies induce similarities in the transitional dynamics of units with different characteristics.

When both n and T are large, there are at least four approaches one can use to estimate the parameters of the model and interesting continuous function of them, which have economic and/or policy interpretations:

1. Estimate the parameters for each unit $i = 1, \ldots, n$ separately using the time series dimension of the panel, (call the estimator α_{iA}); construct the required continuous function (steady state, long run effect, etc.) $h(\alpha_{iA})$ and average to obtain a "typical" effect i.e. $h_A(\alpha) = \frac{1}{n} \sum_i h(\alpha_{iA})$.

2. Pool cross sections and time series, estimate one average parameter vector (call the estimator α_p), and construct one average function $h(\alpha_p)$.

3. Average over n for each $t = 1, \ldots, T$; estimate the parameter vector (call the estimator α_{TS}) and the relevant function $h(\alpha_{TS})$ using the constructed average time series.

4. Average over T and estimate the parameter vector (call the estimator α_{CS}) and the relevant function $h(\alpha_{CS})$ using the constructed average cross sectional data.

Example 8.12 *The magnitude of the savings and investment correlation in open economies (the so-called Feldstein and Horioka puzzle) has attracted the attention of several researchers. Here a large cross section of data for national savings and domestic investments is typically available so all four estimators are feasible. Nevertheless, the literature has concentrated on the average cross sectional estimator and regressions of the type* $(\frac{Sa}{GDP})_i = \varrho_i + A(\frac{Inv}{GDP})_i + e_i$ *are run where* $(\frac{Sa}{GDP})_i$ *is the average saving rate and* $(\frac{Inv}{GDP})_i$ *is the average investment rate for unit i over the sample. Since both saving and investment rates are correlated over time, and since the sample typically includes both OECD and LDC countries, one may guess that a reasonable empirical model could be* $(\frac{Sa}{GDP})_{it} = \varrho_i + \alpha_{1i}(\frac{Sa}{GDP})_{it-1} + \alpha_{2i}(\frac{Inv}{GDP})_{it} + \alpha_{3i}(\frac{Inv}{GDP})_{it-1} + e_{it}$*. Hence, one may be interested in knowing how* α_{CS} *relates to* $\alpha_{ji}, j = 1, 2, 3$ *and whether systematic biases are present.*

The task of this section is to analyze the properties of the four estimators when dynamic heterogeneity is suspected to exist and highlight the problems one is likely to encounter in practical situations. To anticipate the results the first estimator is consistent and a modified version of the fourth can also yield consistent estimates of $h(\alpha)$ (but not necessarily of α) when $T \to \infty$. However, in general, α_p and α_{TS} are inconsistent for $T \to \infty$, producing inconsistent estimates of $h(\alpha)$.

The model we consider has the form:

$$y_{it} = A_{1i}(\ell) y_{it-1} + A_{2i}(\ell) x_{it} + \varrho_i + e_{it} \tag{8.19}$$

$$\alpha_{ji} = \alpha_j + v_{ji} \quad j = 1, 2 \tag{8.20}$$

where $\alpha_{ji} = vec(A_{ji}(\ell))'$. Given (8.19)-(8.20), possible functions of interest are $h_1(\alpha_i) = E(1 - A_{1i}(1))^{-1} A_{2i}(1)$, the long run effect of changes in x_{it} on y_{it}, $h_2(\alpha_i) = E(1 - A_{1i}(1))^{-1} A_{1i}(1)$, the mean lag effect and $h_3(\alpha_i) = (1 - A_{1i}(1))^{-1}$, the convergence rate, etc., where $\alpha_i = vec(A_{1i}(\ell), A_{2i}(\ell))'$.

Note that while we have specified how $A_{ji}(\ell)$ are distributed across i, we could have also specified how $h(\alpha_i)$ are distributed across i; for example, we could have assumed that

$$h(\alpha_i) \quad = \quad h(\alpha) + v_i^h \tag{8.21}$$

Most of the arguments given below go through also with the specification given in (8.21). To insure that the problem is well defined we make four assumptions:
• x_{it} and $e_{i\tau}$ are mutually independent for all $t,\ \tau$ and independent of $v_i = [v_{1i}, v_{2i}]$ (or of v_i^h) with $e_{it} \sim (0,\ \sigma^2_{e_i})$ and $v_{ji} \sim (0, \sigma^2_{v_j}),\ j = 1, 2$.
• The $m_2 \times 1$ vector x_{it} satisfies $x_{it} = \bar{x}_i + \rho x_{it-1} + e^x_{it}$ where $\bar{x}_i$ is the mean and the eigenvalues of ρ are all less then one in absolute value; $e^x_{it} \sim iid\ (0, \sigma^2_{e^x_i})$ and $\lim_{T\to\infty} \frac{1}{T} \sum_{\tau=1}^{T} ACF(\tau) = 0$ (this condition is referred as mean square ergodicity).
• $|\frac{1}{n} \sum_{i=1}^{n} \bar{x}_i \bar{x}_i'| \neq 0$ for a finite n and $\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \bar{x}_i \bar{x}_i' = \Sigma_{xx}$.
• $|A_{1i}(1)| < 1$. The cross sectional moments of $A_{ji}(\ell)$ and of $h(\alpha_i)$ exist and are finite for each i.

These assumptions imply, among other things, that e_{it} are innovations in y_{it}, that x_{it} are strictly exogenous, that y_{it} is stationary and that $h(\alpha_i)$ is computable.

8.3.1 Average time series estimator

When T is sufficiently large, one can run separate regressions for each i, compute $h(\alpha_{iA})$ and average the results to obtain a "typical" effect. When both n and $T \to \infty$, α_{iA}, $h(\alpha_{iA})$, $h_A(\alpha_i)$ yield consistent estimates of $\alpha_i, h(\alpha_i)$ and of $h(\alpha)$. Clearly if T is short estimates of $A_{1i}(\ell)$ are biased and, unless cointegration is present, estimates of $A_{2i}(\ell)$ will also be biased. These biases, in turn, induce biases and inconsistencies in $h_A(\alpha)$ even when n is large. Intuitively, averaging biased estimates will not, in general, eliminate the bias.

Exercise 8.20 *Consider an heterogeneous AR(1) model with no exogenous variables and unit specific fixed effects. Show that estimates of A_{1i} are biased if T is small and that the mean lag effect $E(1 - A_{1i})^{-1} A_{1i}$ will also be biased, regardless of the size of n.*

To show that $h_A(\alpha)$ is consistent rewrite the model as

$$y_{it} = \varrho_i + X_{it}\alpha_i + e_{it} \tag{8.22}$$

where $X_{it} = [y_{it-1}, \ldots y_{it-q_1}, x_{it}, \ldots x_{it-q_2}]$. Then $\alpha_{iA} = (X_i' \Omega_T X_i)^{-1} (X_i' \Omega_T y_i)$ where X_i is a $T \times (q_1 - 1 + q_2)$ matrix, y_i a $T \times 1$ vector, $\Omega_T = I_T - \mathbf{1}_T (\mathbf{1}_T' \mathbf{1}_T)^{-1} \mathbf{1}_T'$ and $\mathbf{1}_T$ a $T \times 1$ unit vector. Let $\alpha_A = \frac{1}{n} \sum_i \alpha_{iA}$ and let $\bar{\alpha} = \frac{1}{n} \sum_i \alpha_i$.

Exercise 8.21 *Give conditions that guarantee $plim_{T\to\infty} \alpha_A = \bar{\alpha}$ (Hint: note that $plim_{T\to\infty} \alpha_A = \bar{\alpha} + \frac{1}{n} \sum_i p\lim_{T\to\infty} (\frac{(X_i' \Omega_T X_i)}{T})^{-1} p\lim_{T\to\infty} (\frac{X_i' \Omega_T e_i}{T})$).*

If also $n \to \infty$, $\bar{\alpha} \to E(\alpha)$ since α_i are iid across i so that α_A is consistent and an estimate of the covariance matrix of α_A is $\Sigma_\alpha = \frac{1}{n(n-1)} \sum_i (\alpha_{iA} - \alpha_A)(\alpha_{iA} - \alpha_A)'$.

Exercise 8.22 *Show that* $E(\Sigma_\alpha) = (1-\frac{1}{n})\sum_i \Sigma_{\alpha_i} + \sum_i E(\alpha_{iA})E(\alpha'_{iA}) - \frac{1}{n}\sum_i \sum_{i'} E(\alpha_{iA})E(\alpha'_{i'A})$ *where the last two terms measure small sample biases. Argue that for* n *fixed, if* T *is large, the bias disappears and* Σ_α *is consistent (You need a lot of algebra to show this!)*

It is immediate to show that any of the $h(\alpha)$ we consider is also consistent. For example, h_{2A} converges to $E(\frac{A_{2i}(1)}{1-A_{1i}(1)})$ if n is large provided that the expression in the denominator is not zero and its variance is $\Sigma_{h_2(\alpha)} = \frac{1}{n(n-1)}\sum_i (h_2(\alpha_{iA}) - h_{2A}(\alpha))(h_2(\alpha_{iA}) - h_{2A}(\alpha))'$.

Example 8.13 *Suppose we are interested in estimating inflation persistence in G-7 countries where persistence is measured either by the spectral density at frequency zero or by the sum of the coefficients of a regression of inflation on its lags. In the first case, we compute the ACF for inflation in each country and the spectral density at frequency zero is obtained summing up 40 covariances and averaging over the seven countries. In the second case, regressions are performed with 10 lags for each country, the sum of coefficients is computed and an average is taken. We find that the range of* $\mathcal{S}_i(\omega = 0)$ *across* i *is large, that the average persistence is 7.03 and that its cross sectional variance is 3.57. The sum of coefficients is also somewhat dispersed: on average, it equals 1.32 and its variance is 0.42. Hence, both statistics suggest that inflation is indeed persistent.*

Exercise 8.23 *Suppose heterogeneity is of binary form, i.e. there are two groups in the data and their composition is known. Describe how to implement an average estimator for inflation persistence in this case. What kind of properties will the estimator have? Under what conditions will it be consistent?*

Example 8.14 *It is relatively easy to design a test for the hypothesis* $\sigma_i = \sigma$, $\forall i$ *assuming that both* ϱ_i *and* α_i *are heterogeneous. In fact, the estimated residuals of the model are* $e_{it} = y_{it} - \varrho_{iA} + X_{it}\alpha_{iA}$. *Under the null* $\sigma_A^2 = \frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T} e_{it}^2$. *Under the alternative* $\sigma_{iA}^2 = \frac{1}{T}\sum_{t=1}^{T} e_{it}^2$. *Then the concentrated values of likelihood under the null and the alternative are* $\mathcal{L}_{re} \propto -\frac{nT}{2}\ln\sigma_A^2$ *and* $\mathcal{L}_{un} \propto -\frac{T}{2}\sum_{i=1}^{n}\ln\hat{\sigma}_{Ai}^2$ *and* $2(\mathcal{L}_{un} - \mathcal{L}_{re}) \sim \chi^2(n-1)$ *as* $T \to \infty$.

Exercise 8.24 *Propose a LR test for the hypothesis* $\alpha_i = \alpha$ *and* $\sigma_i = \sigma$, *for all* i.

While averaging time series estimates is feasible when T is long enough, e.g. convergence regressions at US state level, studies focusing on the cross state effect of local fiscal policy or cross sectional analyses of unemployment rates and labor accidents, it is relatively unusual to see researchers estimating n separate regressions and averaging the results. A typical alternative is to pool cross sections and time series and directly estimate an average α.

8.3.2 Pooled estimator

Substituting (8.20) into (8.19) we have

$$y_{it} = A_1(\ell)y_{it-1} + A_2(\ell)x_{it} + \varrho_i + e_{it}^p \tag{8.23}$$

$$e_{it}^p = e_{it} + v_{1i}'y_{it-1} + v_{2i}'x_{it} \tag{8.24}$$

Since, e_{it}^p is correlated with both y_{it-1} and x_{it}. OLS estimates of $A_1(\ell)$ and $A_2(\ell)$ in (8.23) are inconsistent. Formal evidence of this fact is provided in the next exercise.

Exercise 8.25 *Suppose $A_{1i}(\ell) = A_{1i}, A_{2i}(\ell) = A_{2i}$ and there is no unit specific intercept. (i) Show that $E(x_{it}, e_{it}^p) = \sum_{\tau=0}^{\infty} E(v_{1i}A_{2i}A_{1i}^{\tau})ACF_i(|\tau+1|)$ where $ACF_i(\tau)$ is the autocovariance of x_{it} at lag τ. Note that this expectation goes to zero if x_{it} is serially uncorrelated. (ii) Show that $E(y_{it-1}, e_{it}^p) = \sum_{\tau=0}^{\infty}\sum_{\tau'=0}^{\infty} E(v_{1i}A_{2i}^2A_i^{\tau+\tau'})ACF_i(|\tau-\tau'|)+\sigma_i^2\sum_{\tau=1}^{\infty} E(v_{1i}A_{1i}^{2\tau})+\sum_{\tau=0}^{\infty} E(v_{2i}A_{2i}A_i^{\tau})ACF_i(|\tau+1|)$. Argue that this term does not vanish even when x_{it} is iid.*

When dynamics are heterogeneous and the data pooled, a standard instrumental variable approach is unlikely to work. In fact, given the structure of e_{it}^p, it is difficult to find instruments which are correlated with the regressors and, at the same time, uncorrelated with the error.

Example 8.15 *Consider an AR(1) version of the model and let $z_{it} = [x_{it-1}, \dots x_{it-\tau}]$ be a vector of instruments. z_{it} is a potential candidate since it is uncorrelated with the e_{it} and it is correlated with the regressors of (8.23). However, solving y_{it} from (8.23) we have that*

$$E(v_{it}z_{it}) = E(\frac{\varrho_i v_{1i}}{1-A_{1i}})E(z_{it}) + \sum_{\tau=0}^{\infty} E(v_{1i}A_{1i}^j A_{2i}')E(x_{it-j-1}z_{it}) \tag{8.25}$$

$$E(y_{it-1}z_{it}) = E(\frac{\varrho_i}{1-A_{1i}})E(z_{it}) + \sum_{\tau=0}^{\infty} E(A_{1i}^j A_{2i}')E(x_{it-j-1}z_{it}) \tag{8.26}$$

For z_{it} to be a valid set of instruments we must have that $E(v_{it}z_{it}) = 0$ and $E(y_{it-1}z_{it}) \neq 0$. Staring at (8.25) and (8.26) it is clear that the two sets of conditions cannot be simultaneously satisfied since, in general, $E(\frac{\varrho_i v_{1i}}{1-A_{1i}}) \neq 0$ or $E(v_{1i}A_{1i}^j A_{2i}') \neq 0$.

Exercise 8.26 *Show that in the setup of example 8.15, z_{it} is a valid instrument set if $A_{1i} = A_1, \ \forall i$.*

Since pooled estimators are widely used, it is worthwhile to study the type of biases and inconsistencies they produce when dynamic heterogeneity is present. In what follows we focus on the simplest version of the model (8.19)-(8.20), where there is one lag of y_{it} and one exogenous variable.

Exercise 8.27 *Let $x_{it} = \bar{x}_i + \rho_x x_{it-1} + e_{it}^x$ where $\rho_x < 1$ and $e_{it}^x \sim iid(0, \sigma_{x_i}^2)$; let $\sigma^2 = \frac{1}{n}\sum_i \sigma_i^2$ and $\sigma_x^2 = \frac{1}{n}\sum_i \sigma_{x_i}^2$. Assume that $A_{1i} = A_1, \ \forall i$ but that A_{2i} differs across i. (i) Show that $p\lim_{n,T\to\infty} A_{1p} = A_1 + \frac{\rho_x(1-A_1\rho_x)(1-A_1^2)\sigma_{A_2}^2}{\psi_1}$ and that $p\lim_{n,T\to\infty} A_{2p} = A_2 - \frac{A_2\rho_x^2(1-A_1^2)\sigma_{A_2}^2}{\psi_1}$ where $\psi_1 = \frac{\sigma^2}{\sigma_x^2}(1-\rho_x^2)(1-A_1\rho_x)^2+(1-A_1^2\rho_x^2)\sigma_{A_2}^2+(1-\rho_x^2)A_2$ and $\sigma_{A_2}^2 = var(A_2)$. (ii) Show that the large sample bias of A_{1p} is positive when $\rho_x > 0$ while $p\lim \hat{A}_2 < A_2$ for all parameter values. Argue that the larger is the degree of heterogeneity (i.e. the larger is $\sigma_{A_2}^2$), the greater will be the bias. Show that the bias disappears if and only if $\rho_x = 0$.*

(iii) Show that $p\lim_{n,T\to\infty} \frac{A_{2p}}{1-A_{1p}} = \frac{A_2}{(1-A_1)(1-\rho_x)\psi_2}$ *with* $\psi_2 = \frac{(1+A_1)\sigma^2_{A_2}}{\frac{\sigma^2}{\sigma^2_x}(1+\rho_x)(1-A_1\rho_x)^2+(1+\rho_x)(A_2^2+\sigma^2_{A_2})} > 0$.

Exercise 8.27 shows that OLS overestimates both A_1 and $\frac{A_2}{1-A_1}$ when $\rho_x > 0$ and that the bias washes out if either $\rho_x = 0$ or $\sigma^2_{A_2} = 0$. Furthermore, it is easy to see that if $\rho_x \to 1$, $p\lim A_{1p} = 1$ and $p\lim A_{2p} = 0$ irrespectively of the true value of A_1. Finally, when $A_1 \to 1$, $p\lim A_{1p} = A_1$ and $p\lim A_{2p} = A_2$ (this is not necessarily true if $A_1 = 1$).

The results of exercise 8.27 appear to depend on the presence of serial correlation in the exogenous variables. However, $\rho_x \neq 0$ is inessential and a similar result obtains when x_{it} are iid but current and lagged values of the x_t's enter the regression, as shown next.

Example 8.16 *Suppose that the true model is* $y_{it} = \varrho_i + A_{i2}x_{it} + A_{i3}x_{it-1} + e_{it}$ *and that* $A_i = [A_{i2}, A_{i3}] = A + v_i$, $v_i \sim (0, \Sigma_v)$. *Suppose that* x_{it} *are iid and suppose an investigator estimates* $y_{it} = a_{i1}y_{it-1} + a_{i2}x_{it} + a_{i3}x_{it-1} + \varrho_i + \epsilon_{it}$. *Using OLS on the pooled model and letting both* T *and* n *go to infinity we have that* $a_{2p} = A_2$ *and that* $a_{3p} = A_3 - a_1^*A_2$ *where* $a_1^* = p\lim a_{1p} = \frac{\sigma_{12}}{\sigma_{11}+\sigma_{22}+(A_3^2+\frac{\sigma^2}{\sigma^2_x})}$ *where* σ_{ij} *are the elements of* Σ_v. *Therefore, no matter how large* T *and* n *are,* $a_1^* = 0$ *if and only if* $\sigma_{12} = 0$. *Hence estimates of the long run effect of* x *on* y, $\frac{a_{2p}+a_{3p}}{1-a_{1p}}$ *will converge to* $A_2 + A_3 + \frac{a_1^*A_3}{1-a_1^*} \neq A_2 + A_3$ *as* $T \to \infty$.

The biases and inconsistencies of example 8.16 occur because heterogeneity in the coefficients of the x's is ignored. Note also that serial correlation in the x's makes the problem worse. Clearly, the size of the bias depends on the sign and the magnitude of σ_{12} and ρ_x. When these are positive, $a_1^* > 0$ and there will be a tendency to underestimate the impact of x_{it-j} on y_{it} and overestimate its long run effect.

When T is short, a pooled estimator is inconsistent even without dynamic heterogeneity. As we have seen, in this situation it is typical to use either an Anderson-Hsiao (AH) or a GMM estimator after differencing. But while with homogeneous dynamics both approaches produce consistent estimates of the parameters, this is not the case in the current setup.

Example 8.17 *After first differencing (8.23) and (8.24), become*

$$\Delta y_{it} = A_1(\ell)\Delta y_{it-1} + A_2(\ell)\Delta x_{it} + \Delta e^p_{it} \quad (8.27)$$

$$\Delta e^p_{it} = \Delta e_{it} + v'_{1i}\Delta y_{it-1} + v'_{2i}\Delta x_{it} \quad (8.28)$$

Clearly, any instrument uncorrelated with the errors will also be uncorrelated with the regressors. For example, lagged values of y_{it} *are not valid instruments since they depend on* v_{ji} *which, in turn, are correlated with* Δe^p_t. *Similarly, current and lagged values of* x_{it} *are not valid instruments even when they are uncorrelated with* e_{it}.

There is one special case when differencing solves the problems. In fact, when $v_{i1} = 0\ \forall i$, simple algebra shows that $E(\Delta e^p_{it}z_{it}) = 0$ but $E(\Delta y_{it-1}z_{it}) = \sum_{\tau=0}^{\infty} E(A^\tau_{1i}A'_{2i})E(\Delta x_{it-\tau-1}z_{it}) \neq 0$. Therefore IV estimation after differencing may yield consistent estimators of the mean of A_{1i} and A_{2i} if lags of x_{it} are used (the lags of y_{it} are invalid instruments).

Exercise 8.28 *Show that, if A_{1i} is independent of A_{2i}, differencing and using appropriate lags of Δx_{it} as instruments yields consistent estimates of the mean value A_1 and A_2.*

Example 8.18 *(Sorensen, Wu, Yosha) Suppose we wish to examine the cyclicality of government expenditure over a sample of countries and run the regression $(\frac{G}{GDP})_{it} = \varrho_i + \alpha_{1i}\Delta GDP_{it} + \alpha_{2i}\Delta GDP_{it-1} + e_{it}$. If $\alpha_{ji} = \alpha_j \ \forall i$, a pooled regression after differencing produces consistent estimates of the responses of $(\frac{G}{GDP})_{it}$ to a shock in ΔGDP_{it} for each i if proper instruments are used. Consistent estimates could also be obtained even if dynamic heterogeneity is neglected as long as $cov(\alpha_{1i}, \alpha_{2i}) = 0$. Furthermore consistent estimates could be obtained even when $(\frac{G}{GDP})_{it-1}$ enters the regression as long as its coefficient is homogeneous across i.*

8.3.3 Aggregate time series estimator

Let $\bar{y}_t = \frac{1}{n}\sum_i y_{it}$, $\bar{x}_t = \frac{1}{n}\sum_i x_{it}$, $\bar{e}_t = \frac{1}{n}\sum_i e_{it}$, $\bar{\varrho} = \frac{1}{n}\sum_i \varrho_i$. The model we consider is:

$$\bar{y}_t = A_1(\ell)\bar{y}_{t-1} + A_2(\ell)\bar{x}_t + \bar{\varrho} + \bar{e}_t^{TS} \tag{8.29}$$

$$\bar{e}_t^{TS} = \bar{e}_t + \frac{1}{n}\sum_{i=1}^{N}(v_{1i}'y_{it-1} + v_{2i}'x_{it}) \tag{8.30}$$

Serial correlation in x_{it} clearly produces a complex serial correlation pattern in $\bar{e}_t^{TS}$. What is perhaps less immediate to see is that $\bar{e}_t^{TS}$ is correlated with the regressors so that OLS applied to (8.29) will yields inconsistent estimates even when T or $T, n \to \infty$.

Example 8.19 *We demonstrate this problem when there is only one lag of y_{it}, when $dim(x_{it}) = 1$ and $A_2(\ell) = 0, \ \forall \ell > 0$. The OLS estimator of A_2 in (8.29) is $A_{2,TS} - A_2$ $= \frac{(\sum_t \bar{y}_{t-1}^2)(\sum_t \bar{x}_t\bar{e}_t^{TS}) - (\sum_t \bar{x}_t\bar{y}_{t-1})(\sum_t \bar{y}_{t-1}\bar{e}_t^{TS})}{(\sum_t \bar{y}_{t-1}^2)(\sum_t \bar{x}_t^2) - (\sum_t \bar{x}_t\bar{y}_{t-1})^2}$. Then, as $T \to \infty$,*

$$p\lim A_{2,TS} - A_2 = \frac{(E\bar{y}_{t-1}^2)(E\bar{x}_t\bar{e}_t^{TS}) - (E\bar{x}_t\bar{y}_{t-1})(E\bar{y}_{t-1}\bar{e}_t^{TS})}{(E\bar{y}_{t-1}^2)(E\bar{x}_t^2) - (E\bar{x}_t\bar{y}_{t-1})^2} \tag{8.31}$$

For consistency we need $\sum_t \bar{x}_t\bar{e}_t^{TS} \to 0$ and $\sum_t \bar{y}_{t-1}\bar{e}_t^{TS} \to 0$. But

$$E(\bar{x}_{t-\tau}\bar{e}_t^{TS}) = \frac{1}{n}\sum_{i=1}^{N}\sum_{\tau'=1}^{\infty} E(v_{2i}A_{2i}A_{1i}^{\tau'})E(\bar{x}_{t-\tau}, x_{i,t-\tau'-1}) \tag{8.32}$$

For $E(\bar{x}_{t-\tau}\bar{e}_t^{TS}) = 0$, we need that either x_t are serially uncorrelated (the second term vanishes) or that there is no parameter heterogeneity (the first term vanishes). Hence, the expression in the numerator (8.31) will not, in general, be equal to zero.

Exercise 8.29 *Show that, in general, $\sum_t \bar{y}_{t-1}\bar{e}_t^{TS}$ does not converge to zero as $T \to \infty$. Show the conditions under which this may occur.*

Since the terms in (8.32) are of order n^{-1} and since $E(\bar{x}_t e_t^{\bar{T}S})$ and $E(\bar{y}_{t-1} e_t^{\bar{T}S})$ converge to a finite limit, increasing n will not eliminate the inconsistency of the aggregate time series estimator. Also, since the serial correlation properties of $\bar{e}^{TS}$ are sufficiently complex, an IV approach is unlikely to work. For example, staring at (8.32), one can see that lags of $\bar{x}_t$ are invalid instruments. The problem is similar to the one encountered with the pooled estimator: variables which are uncorrelated with $\bar{e}_t^{TS}$ will be also uncorrelated with the regressors. Therefore valid instruments are hard to find.

One reason for why the aggregate time series estimator is inconsistent is that averaging over n does aggregate cross sectional information optimally. Pesaran (1995) showed that in an heterogeneous dynamic model the optimal cross sectional aggregator has the form $\bar{y}_t = \sum_{\tau=0}^{\infty} a_\tau' \bar{x}_{t-\tau} + \bar{\epsilon}_t$, where $a_\tau = E(A_{2i} A_{1i}^\tau)$, $\tau = 0, 1, \ldots$ and $\bar{\epsilon}_t$ are iid independent of x_t. (8.29) misspecifies this expression because - important regressors, correlated with included ones, are omitted from the specification. Therefore, inconsistencies are produced.

Exercise 8.30 *Show that if A_{1i} and A_{2i} are independently distributed, there is only one lag of y_{it} in the model, consistent estimates of $\frac{A_2}{1-A_1}$, can be obtained from an aggregate time series specification using an infinite distributed lag regression of $\bar{y}$ on $\bar{x}$.*

Exercise 8.31 *Consider the case where $h(\alpha_i)$ is distributed as in (8.21). Show that the aggregate time series estimator of $h(\alpha)$ is inconsistent*

Example 8.20 *Continuing with example 8.13 we compute the spectral density at frequency zero of inflation using pooled and aggregate time series estimators. The point estimate for pooled data is 9.84, which is within one standard error of the estimate obtained with the average estimator. With aggregated data the point estimate is 13.00, a value which is in the 99 percentile of the distribution of the average estimator. The point estimate of the sum of coefficients of the regression is 0.91 in the pooled case and 0.97 in the aggregate case, both of which are substantially smaller than the point estimate obtained with the average estimator, but not significantly different from it.*

8.3.4 Average Cross sectional Estimator

The average cross sectional estimator is also popular in applied work and believed to unbiasedly measure both the parameters and interesting continuous functions of them. But while in static models such a presumption is correct, this is not necessarily the case when heterogeneous dynamic models are considered.

Example 8.21 *(Fatas and Mihov) There has been some interest in examining whether macroeconomic volatility, typically measured by the standard deviation of output growth, is systematically related to government size, typically measured by the log of the share of government expenditure to GDP, since simple Keynesian models predict a negative relationships between the two. To check this hypothesis, the literature has estimated one volatility for each unit, averaged the expenditure share over t and run a cross sectional regression,*

with or without additional controls. Typically, a negative coefficient is found but there may be doubts about the reliability of estimates since, as we will see, neglecting dynamic heterogeneity induces negative and large biases.

Let $\bar{y}_i = \frac{1}{T}\sum_t y_{it}$, $\bar{x}_i = \frac{1}{T}\sum_t x_{it}$, $\bar{e}_i = \frac{1}{T}\sum_t e_{it}$. Then the model we are estimating is:

$$\begin{aligned} \bar{y}_i &= A_1(\ell)\bar{y}_{i,-1} + A_2(\ell)\bar{x}_i + \varrho_i + \bar{e}_i^{CS} && (8.33)\\ \tilde{e}_i^{CS} &= e_i + v_{1i}'\bar{y}_{i,-1} + v_{2i}'\bar{x}_i && (8.34) \end{aligned}$$

The regression defined by (8.33) is the so-called "between" regression of the dynamic model. Such a model when estimated with OLS yields inconsistent estimates of $A_j(\ell)$ when $n \to \infty$ or $n, T \to \infty$ since $\bar{e}^{CS}$ is correlated with the regressors, even when n is large. Hence, functions of a "between" regression estimator obtained from an heterogenous panel will also be inconsistent.

One way to produce consistent estimates is to replace (8.33) with:

$$\bar{y}_i = A_{2i}(\ell)\bar{x}_i + A_{1i}(\ell)(\bar{y}_i - \Delta_T y_i) + \bar{e}_i \quad (8.35)$$

where $\Delta_T y_i = \frac{(y_{iT} - y_{i0})}{T}$. Notice that equation (8.35) is equivalent to $\bar{y}_i = (1 - A_{1i}(\ell))^{-1}A_{2i}\bar{x}_i - (1 - A_{1i}(\ell))^{-1}A_{1i}(\ell)\Delta_T y_i + (1 - A_{1i}(\ell))^{-1}\bar{e}_i \equiv a_{1i}\bar{x}_i + a_{2i}\Delta_T y_i + \bar{\epsilon}_i$. If $a_{ji},\ j = 1, 2$ are randomly distributed around the mean, then

$$\begin{aligned} \bar{y}_i &= a_1\bar{x}_i - a_2\Delta_T y_i + \bar{\epsilon}_i^{cs} && (8.36)\\ \bar{\epsilon}_i^{cs} &= (1 - A_{1i}(\ell))^{-1}\bar{e}_i + v_{2i}'\bar{x}_i - v_{1i}'\Delta_T y_i && (8.37) \end{aligned}$$

In the next exercise we ask the reader to verify that consistent estimates of a_1, a_2 can be obtained with OLS if T is large enough.

Exercise 8.32 *Let the cross section estimator of a_1 be $a_{1,cs}' = (\sum_i \bar{x}_i\bar{x}_i)^{-1}((\sum_i \bar{x}_i\bar{y}_i')$. (i) Show that $E[a_{1,cs}' - a_1'] = (\sum_i \bar{x}_i\bar{x}_i')^{-1}(\sum_i (\bar{x}_i\bar{x}_i')v_{2i} + (\sum_i \bar{x}_i\bar{x}_i')^{-1}\sum_i (1 + a_{2i})\bar{x}_i\bar{e}_i^{cs} - (\sum_i \bar{x}_i\bar{x}_i')^{-1}$ or using the assumptions made that $E[a_{1,cs}' - a_1'] = -\sum_{\tau=0}^{\infty}(\sum_i \bar{x}_i\bar{x}_i')^{-1}\sum_{i=1}^{n}\bar{x}_i (\bar{x}_{i,-\tau} - \bar{x}_{i,-\tau-1})'E(a_{2i}A_{2i}A_{1i}^{\tau})$. Hence, for finite T, $E(a_{1,cs}')$ will be biased even if $n \to \infty$. iii) Show that for $T \to \infty$ $\bar{x}_i(\bar{x}_{i,-\tau} - \bar{x}_{i,-\tau-1})' = O_p(T^{-1})$. Conclude that since $E(a_{2i}A_{2i}A_{1i}^{\tau})$ is finite, $E[a_{1,cs}' - a_1'] \to 0$ in probability.*

It is important to stress that the estimator in (8.33) is inconsistent not because we have misspecified the model. The omitted term, $\Delta_T y_i$, is asymptotically uncorrelated with the level variables so that it will not affect estimates of long run effects. The inconsistency is instead produced by the correlation between the error and the regressors.

Exercise 8.33 *Using (8.35) as reference, one lag of y and only contemporaneous x's, show that if A_{2i} and A_{1i} are random (instead of $a_{ji}, j = 1, 2$) $a_{2,cs}$ is consistent.*

In sum, if dynamic heterogeneity is present, cross sectional regressions where variables are time series averages for each unit are problematic since, for fixed T, these will be inconsistent, even when n is large.

In certain applied situations one may want to use cointegration ideas to get estimates of the parameters of a dynamic model. If x_{it} were integrated variables and each i had its own cointegrating relationship, then unit specific regressions yield superconsistent estimates of $A_2(1)$ and of $h(\alpha_i)$. Then the average estimator of $h_A(\alpha)$ will also be consistent. Note that since parameter estimates in this case converge at the rate T and since the average converges at the rate $\sqrt{n}$, the estimator of $h(\alpha)$ converges at the rate $T\sqrt{n}$. Note also that a pooled regression will not yield a consistent estimate of α even in the presence of cointegration. This is because the error term has an I(0) component, the residuals of the cointegrating relationship for unit i, and an I(1) component, the product of the difference between the coefficient of each i from the imposed common coefficient and the I(1) regressor. Therefore the composite error is I(1) and the regression does not define a cointegrating relationship.

The next exercise examines what happens to the other two estimators when the variables of the model are integrated.

Exercise 8.34 *(i) Consider the aggregate time series estimator applied to the model* $\bar{y}_t = A_2\bar{x}_t + \bar{e}_t^{TS}$ *where* $\bar{e}_t^{TS} = \bar{e}_t + \frac{1}{n}\sum_i v_i x_{it}$, $A_{2i} = \bar{A}_2 + v_i$, *and* x_{it} *strictly exogenous. Show that if* $x_{i,t} = x_{i,t-1} + e_{i,t}^x$, $e_{i,t}^x \underset{P}{\sim} (0, \sigma_{x_i}^2)$, $A_{2,TS}$ *is inconsistent.*
(ii) Show that $A_{2,CS} = A_2 + \frac{\sum_i v_i \bar{x}_i^2 + \bar{e}_i \bar{x}_i}{\sum_i \bar{x}_i^2}$. *Show that if* $x_{it} = x_{it-1} + \bar{x}_i + e_{it}^x$, $e_{it}^x \sim (0, \sigma_{x_i}^2)$, $A_{2,CS}$ *is consistent for* T *fixed and* $n \to \infty$.

8.3.5 Testing for dynamic heterogeneity

Since the presence of heterogeneous dynamics causes problems to standard estimators even when first differencing and instrumental variables are used, it is crucial to have a way to assess whether homogeneity holds in the sample under consideration. One way of testing for dynamic heterogeneities is to use a Hausman-type test, which we described in chapter 5. The idea of the test is very similar to the one presented in section 8.2: we wish to find one estimator which is consistent under the two hypotheses and another which is consistent (and efficient) under the null and inconsistent under the alternative.

Given these requirements we can compare, e.g., the pooled estimator and the average time series estimator. In fact, under the null of homogeneity both are consistent and the pooled estimator is more efficient since it uses all the available information. Under the alternative of heterogeneity, only the average time series estimator is consistent.

The asymptotic variances of the two estimators are (for fixed n and large T) $\sigma^2 \times (\sum_i plim_{T\to\infty}(X_i'\Omega_T X_i/T))^{-1}$ and $\frac{\sigma^2}{n^2}\sum_i (plim_{T\to\infty}(X_i'\Omega_T X_i/T)^{-1})$. Hence, the covariance matrix of the difference between the estimators is $\frac{\sigma^2}{n}[\frac{1}{n}\sum_i (plim_{T\to\infty}(X_i'\Omega_T X_i/T)^{-1}) - (\sum_i plim_{T\to\infty}(X_i'\Omega_T X_i/nT))^{-1}]$ which is positive definite except when $plim_{T\to\infty}(X_i'\Omega_T X_i/T)) = plim_{T\to\infty}(X_j'\Omega_T X_{i'}/T))$, $i' \neq i$. Then a test for heterogeneity can be conducted using $S_1 = \hat{\sigma}_A^2(\alpha_A - \alpha_P)'\Sigma^{-1}(\alpha_A - \alpha_P)$ where $\hat{\sigma}_A^2 = \frac{1}{n}\sum_i \hat{\sigma}_i^2$; $\Sigma = \frac{1}{n^2}\sum_i (X_i'\Omega_T X_i)^{-1} -$

$(\sum_i X_i'\Omega_T X_i)^{-1}$ and $\alpha = vec(A_{1i}(\ell), A_{2i}(\ell))$. Under the null that $A_{1i}(\ell) = A_1(\ell),\ A_{2i}(\ell) = A_2(\ell),\ \sigma_i^2 = \sigma^2,\ \forall i$ $\mathsf{S}_1 \sim \chi^2(dim(x_t) + dim(y_t))$.

Exercise 8.35 *Show that substituting $\hat{\sigma}_p^2$ for $\hat{\sigma}_A^2$ does not change the asymptotic distribution of the test.*

Example 8.22 *The test can also be undertaken on the relevant functions of the parameters. Consider estimating $h_2(\alpha) = E(1 - A_{1i}(1))^{-1}A_{2i}(1))$. The pooled estimator is $h_{2P} = (1-A_{1P}(1))^{-1}A_{2P}(1) = E(1-A_{1i}(1))^{-1}A_{2i}(1)) + \frac{(A_{1P}(1)-A_1)E(1-A_{1i}(1))^{-1}A_{2i}(1))+(A_{2P}(1)-A_2)}{1-A_{1P}(1)}$ and the average time series estimator is $h_{2A} = \frac{1}{n}\sum_i h_2(\alpha_i)$. The asymptotic variance of the first estimator is $\frac{\sigma^2}{(1-A_1(1))^2}D(\sum_i p\lim(X_i'\Omega_T X_i/T))^{-1}D'$ and of the second is $\frac{\sigma^2}{n^2(1-A_1(1))^2}D\times$ $(\sum_i (p\lim(X_i'\Omega_T X_i/T)^{-1}))D'$ where $D = (a_2, I_{m_2})$ and I_{m_2} is a $m_2 \times (m_2 + 1)$ matrix.*

Then a test of homogeneity can be based on $\mathsf{S}_2 = \frac{\hat{\sigma}^2}{(1-\hat{A}_1(1))^2}(h_{2A} - h_{2P})'(\hat{D}\hat{\Sigma}\hat{D}')^{-1}$ $(h_{2A} - h_{2P})$ where hat-variables can be obtained from either the pooled or the average time series estimator and $\hat{\Sigma} = \frac{\hat{\sigma}^2}{n^2(1-\hat{A}_1(1))^2}\hat{D}(\sum_i (p\lim(X_i'\Omega_T X_i/n)^{-1}))\hat{D}' - \frac{\hat{\sigma}^2}{(1-\hat{A}_1(1))^2}\hat{D} \times$ $(\sum_i p\lim(X_i'\Omega_T X_i/nT))^{-1}\hat{D}'$. Under the null, $\mathsf{S}_2 \sim \chi^2(m_2)$. Note that $\hat{\Sigma}$ may not be positive definite in small samples.

These tests are appropriate when T is large, since for small T the average time series estimator is biased. When T is small it is still possible to conduct an homogeneity tests for $h(\alpha)$ using the modified cross sectional estimator which is consistent under the alternative. However, since Hausman test is asymptotically justified only when T is large, care must be exercised when comparing the properties of pooled and cross section estimators. It is also worth mentioning that the Hausman test has poor power properties when outliers are present since the variance of the estimators tends to be very large. Similarly, if the cross sectional data is of uneven quality, the null of homogeneity could be difficult to reject.

Exercise 8.36 *Provide a statistics to test heterogeneities in $h_1(\alpha) = (1 - A_1(\ell))^{-1}A_1$ when T is short.*

Example 8.23 *In general, it is difficult to interpret rejections of the homogeneity hypothesis since heterogeneity may result from a misspecified but homogenous model. When homogeneity is rejected, one typically finds a very large dispersion of estimates, with several economically implausible individual estimates, but the average of the estimates may turn out to be quite sensible. Can this pattern provide information for the likely causes of heterogeneity? Suppose that the estimated model is $y_{it} = \alpha_i x_{1it} + \epsilon_{it}$ where $\epsilon_{it} = x_{2it} + e_{it}$ and $x_{2it} = \theta_{it}x_{1it} + v_{it}$ where x_{2it} are omitted variables linked in a time varying fashion to the regressors x_{1it}. Then it is easy to see that $E(\hat{\alpha}_{it}) = \alpha_i + \theta_{it}$. Consequently, the specification error in $\hat{\alpha}_{it}$ is large and significant if x_{2it} are important for y_{it} and θ_{it} non-negligible. If x_{2it} are common to all i's for all t (e.g., commodity prices) then $\hat{\alpha}_T = \frac{1}{n}\sum_i \hat{\alpha}_{iT}$ will have a systematic bias. However, if x_{2it} are random over the cross section, it is possible that*

$E(\theta_{iT}) = 0$, so that average estimates are more reliable. The same result would occur if x_{2it} are randomly correlated for each i across T. Finally, the structure considered in this example may not only cause heterogeneities but also instabilities in each i since the correlation structure between y_{1t} and x_{2t} evolves over time.

8.4 To Pool or not to Pool?

In many applied exercises, a researcher is interested in examining estimates of, say, long run coefficients, elasticities, or impulse responses over the cross section hoping to infer whether certain individual characteristics (say, labor market regulations or government policies) are responsible for the differences. When T is short comparison is difficult and, at times, uninformative since estimates are biased and the estimation uncertainty is large. One question of interest is therefore whether it is possible to improve single unit estimation of the parameters using cross sectional information. We have already seen that complete pooling is efficient under homogeneity but produces biases and inconsistencies if dynamic heterogeneities are present. Here we are concerned with whether some form of pooling is advisable even under heterogeneity and on how partial pooling could be performed in a simple and tractable way.

The simplest procedure one can use to check whether pooling is appropriate is a preliminary test of equality of the coefficients over the cross section. Suppose the model is

$$y_{it} = X_{it}\alpha_i + e_{it} \qquad e_{it} \sim (0, \sigma^2 * I) \tag{8.38}$$

where X_t includes a vector of ones, exogenous and lagged dependent variables and α_i all the regression parameters of unit i. If the null hypothesis is $\alpha_1 = \alpha_2 = \ldots = \alpha_{n_1} = \alpha, \;\; n_1 \leq n$, a pooled model for n_1 units is just the unpooled model with some (exact) linear restrictions on the parameters.

To verify the null hypothesis it is typical to compare the R^2 of two regressions, one with and one without the restrictions. Given that $F = \frac{(R^2_{re} - R^2_{un})/n-1}{R^2_{un}/(nT-n)}$, where R^2_{re} (R^2_{un}) is the R^2 of the restricted (unrestricted) model, choosing the unpooled model based on the regression R^2 is equivalent to choosing the alternative hypothesis when the F-ratio exceeds 1, which implies a significance level of 50%. Information criteria such as the AIC, also imply preference for the alternative hypothesis if the F-ratio is less than one (see Maddala (1992)). Hence, one should be aware that the significance level used when pretesting is different from the one used in standard hypothesis testing. Note also that the pretest estimator of α_i is discontinuous (it is $\alpha_{i,OLS}$ if F-ratio > 1 and α_{OLS} otherwise), so that its asymptotic distribution is complicated. Furthermore, it is dominated under a quadratic loss function by other estimators, see Judge et al (1985, p.72-80).

Apart from theoretical problems, it is very common in applied work to encounter situations like the one described in example 9.29. Therefore, without exact knowledge of the distribution of the observations across units, surprising results may appear.

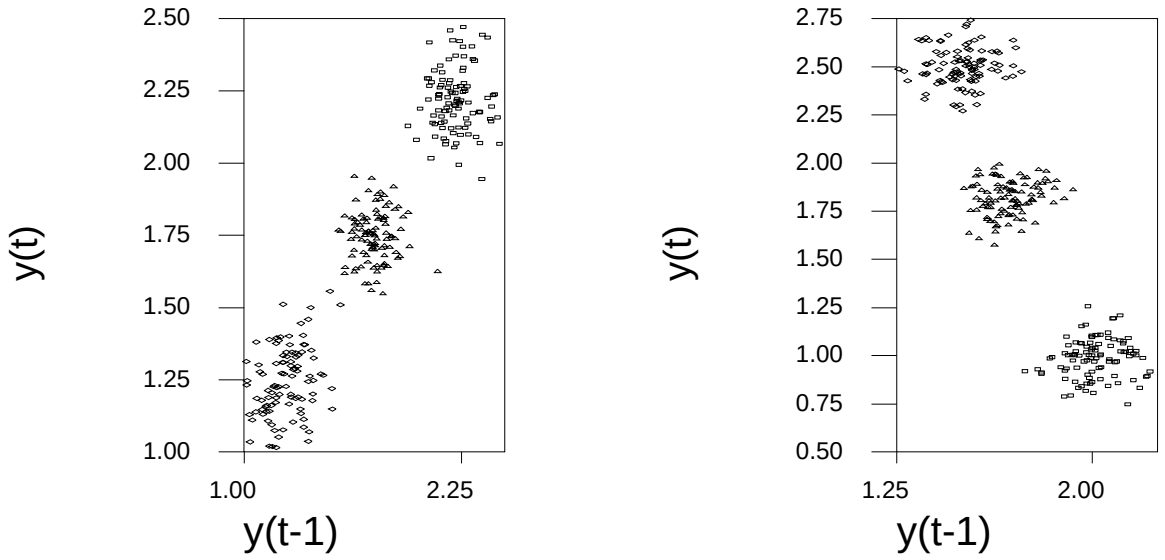


Figure 8.2: Cross sectional distributions

Example 8.24 *Suppose $n = 3$ and that we run a AR(1) regression with no x_t's and unit specific parameters. Figure 8.2 plots a cloud of points for two distributions of i. Here the regression slopes are identical for each i so one may end up pooling observations if the standard error of the intercept is large enough. However, while pooling in the first case will maintain the positive slope (biasing upwards the estimate of the AR coefficient), in the second it will produce negative estimates of the common slope parameters. Since there is no reason to a-priori exclude the second distribution, it is possible that y_{it} and y_{it-1} are positively correlated with individual data and negative correlated with pooled data.*

Given these problems, pretesting does not seem to be the answer to reduce biases and improve standard errors of estimates.

One way to produce improved estimates of the parameters using cross sectional information is a Stein-type shrinkage estimator, i.e.

$$\alpha_{is} = \alpha_p + (1 - \frac{\kappa}{F})(\alpha_{i,ols} - \alpha_p) \quad i = 1, \ldots, n \tag{8.39}$$

Here $\alpha_{i,ols}$ is the OLS estimator obtained with data from unit i, α_p is the pooled estimator, F is the statistics for the null hypothesis $\alpha_i = \alpha \forall i$, (i.e. $F = \frac{(\alpha_i - \alpha)'(\alpha_i - \alpha)}{n\sigma^2}$) and $\kappa = [(n-1)dim(\alpha)-2]/[nT-dim(\alpha)+2]$, which, for large n, reduces to $\kappa \approx dim(\alpha)/(T-dim(\alpha))$.

The Stein-type estimator (8.39) can be formally obtained minimizing the risk of the estimator (see e.g. Judge et al. (1985), p. 83). Note that a Stein-type estimator combines individual and pooled estimates using a weight which, for large n, depends on the dimension

of α relative to the size of time series. The larger is $dim(\alpha)$ relative to T, the smaller will be the shrinkage factor $(1 - \frac{\kappa}{F})$.

Another way to partially pool heterogenous cross sectional information is to use a random coefficient model. Random coefficient models also lead to shrinkage-style estimators but, contrary to Stein estimators, they combine individual estimates with a weighted average of the $\alpha_{i,ols}$. Suppose the model is

$$y_{it} = x_{it}\alpha_i + e_{it} \tag{8.40}$$
$$\alpha_i = \bar{\alpha} + v_i \tag{8.41}$$

where $e_i \sim (0, \sigma_i^2 I)$ and $v_i \sim (0, \Sigma_v)$. There are four approaches one can use to construct improved estimates of α_i which correspond, roughly speaking, to Classical, Bayesian, Prior likelihood and Empirical Bayes approaches.

In a classical approach α and Σ_v are estimable but individual α_i are not. In this case one substitutes (8.41) into (8.40) so that $y_{it} \sim (X_{it}\alpha, \Sigma_{it})$ where $\Sigma_{it} = \sigma_i^2 I + x_{it}'\Sigma_v x_{it}$. This is what the literature typically calls error-component model. The GLS estimator for α, obtained stacking the T observations for each unit is $\alpha_{GLS} = (\sum_i x_i\Sigma_i^{-1}x_i)^{-1}(\sum_i x_i\Sigma_i^{-1}y_i)$ which collapses to the OLS estimator $\alpha_{i,ols} = (\sigma^{-2}x_i'x_i)^{-1}(\sigma^{-2}x_i'y_i)$ if the α_i were fixed.

Exercise 8.37 *Show that* $\alpha_{GLS} = \sum_i (\sum_{j=1}^n \Omega_j^{-1})^{-1}\Omega_i^{-1}\alpha_{i,ols}$ *with* $\Omega_i = (\sigma^2(x_i'x_i)^{-1} + \Sigma_v)$.

Exercise 8.37 shows that the GLS estimator of α is a weighted average of the OLS estimators for each α_i (constructed treating σ_i^2 Σ_v as fixed), with weights given by a function of Σ_v and of the data matrix $(x'x)$. α_{GLS} is unfeasible since Σ_v and σ^2 are unknown. Therefore one would plug $\Sigma_{v,ols} = \frac{1}{n-1}\sum_{i=1}^n(\alpha_{i,ols} - \frac{1}{n}\sum_{i=1}^n \alpha_{i,ols})(\alpha_{i,ols} - \frac{1}{n}\sum_{i=1}^n \alpha_{i,ols})' - \frac{1}{n}\sum_i \sigma_{i,ols}^2(x_i x_i')^{-1}$ and $\sigma_{i,ols}^2 = \frac{1}{T-dim(\alpha_i)}(y_i'y_i - y_i x_i \alpha_{i,ols})^2$ in the GLS formula. Since $\Sigma_{v,ols}$ is not necessarily positive definite, it is typical to neglect the last term of the expression: the resulting estimator is biased but non-negative definite and consistent as $T \to \infty$.

In the other three approaches, equation (8.41) is treated as a prior and α and Σ_v represent a second layer of parameters (the hyperparameters) describing the features of the prior. If α and Σ_v were known, the posterior of α_i is normal with mean $\tilde{\alpha}_i = (\frac{1}{\sigma_i^2}x_i'x_i + \Sigma_v^{-1})^{-1}(\frac{1}{\sigma_i^2}x_i'x_i\alpha_{i,ols} + \Sigma_v^{-1}\bar{\alpha})$ where $\alpha_{i,ols}$ is the OLS estimator of α_i. It is easy to see that if Σ_v is large $\tilde{\alpha}_i \to \alpha_{i,ols}$; that is, there is no information in the prior which can be used to improve estimates of α_i. α_{GLS} is related to $\tilde{\alpha}_i$ via $\alpha_{GLS} = \frac{1}{n}\sum_{i=1}^n \tilde{\alpha}_i$; that is, the GLS estimator equals the sample average (over i) of the Bayesian estimator $\tilde{\alpha}_i$.

When $\bar{\alpha}$, σ_i^2, Σ_v are unknown, one should specify a prior distribution for these parameters, see e.g. chapter 9. In general, no analytical solution for the posterior mean of α_i exists. If normality is likely to hold, one could approximate posterior means with posterior modes (see Smith (1973)), i.e. use

$$\bar{\alpha}^* = \frac{1}{n}\sum_{i=1} \alpha_i^* \tag{8.42}$$

$$(\sigma_i^*)^2 = \frac{1}{T+2}[(y_i - x_i\alpha_i^*)'(y_i - x_i\alpha_i^*)] \tag{8.43}$$

$$\Sigma_v^* = \frac{1}{n - \dim(\alpha) - 1}[\sum_i (\alpha_i^* - \bar{\alpha}^*)(\alpha_i^* - \bar{\alpha}^*) + \kappa \tag{8.44}$$

where, "*" indicates modal estimates and, typically, $\kappa = \text{diag}[0.001]$. Note that the use of modal estimates does not change the form of the estimates of $\tilde{\alpha}_i$ and therefore of $\tilde{\alpha}$.

As an alternative, one can use the so-called prior likelihood approach. Roughly speaking, one jointly selects $(\alpha_i, \sigma_i^2, \bar{\alpha}, \Sigma_v)$ to maximize $-\frac{T}{2}\sum_{i=1}^n \ln \sigma_i^2 - 0.5[\sum_{i=1}^n \frac{1}{\sigma_i^2}(y_i - x_i\alpha_i)'(y_i - x_i\alpha_i) - nln|\Sigma_v| - \sum_{i=1}^n(\alpha_i - \bar{\alpha})^{-1}\Sigma_v^{-1}(\alpha_i - \bar{\alpha})]$. The solution is

$$\alpha_{i,pl} = (\frac{1}{\sigma_{i,pl}^2}x_i'x_i + \Sigma_{v,pl}^{-1})^{-1}(\frac{1}{\sigma_{i,pl}^2}x_i'x_i\alpha_{i,ols} + \Sigma_{v,pl}^{-1}\bar{\alpha}_{pl}) \tag{8.45}$$

$$\bar{\alpha}_{pl} = \frac{1}{n}\sum_{i=1} \alpha_{i,pl} \tag{8.46}$$

$$\sigma_{i,pl}^2 = \frac{1}{T}(y_i - x_i\alpha_{i,pl})'(y_i - x_i\alpha_{i,pl}) \tag{8.47}$$

$$\Sigma_{v,pl} = \frac{1}{n}\sum_{i=1} (\alpha_{i,pl} - \bar{\alpha}_{pl})(\alpha_{i,pl} - \bar{alpha}_{pl})' \tag{8.48}$$

Note the similarities between (8.42)-(8.44) and (8.46)-(8.48).

Exercise 8.38 *Suggest an iterative procedure to obtain* $\alpha_{i,pl}, \bar{\alpha}_{pl}, \sigma_{i,pl}^2, \Sigma_{v,pl}$.

Finally, one could use empirical Bayes (EB) methods. As we will see in more details in the next chapter, this approach treats $\Sigma_v, \sigma_i^2, \bar{\alpha}$, as unknown and estimates them using the predictive density of y in a training sample. An EB estimator is (see e.g. Rao (1975)):

$$\bar{\alpha}_{EB} = \frac{1}{n}\sum_{i=1} \alpha_{i,ols} \tag{8.49}$$

$$\sigma_{i,EB}^2 = \frac{1}{T - dim(\alpha)}(y_i'y_i - y_i'x_i\alpha_{i,ols}) \tag{8.50}$$

$$\hat{\Sigma}_{v,EB} = \frac{1}{n-1}\sum_{i=1} (\alpha_{i,ols} - \bar{\alpha}_{EB})(\alpha_i - \bar{\alpha}_{EB})' - \frac{1}{n}\sum_{i=1} (x_i'x_i)^{-1}\sigma_i^2 \tag{8.51}$$

Clearly, the fully Bayes and the Empirical Bayes estimators of α are similar but while the former is an average of $\tilde{\alpha}_i$, the latter is an average of OLS estimates. Note that both estimators can be computed in two steps and do not require iterative solutions. Alternative Empirical Bayes estimators for dynamic heterogeneous panels are presented in chapter 10.

Pooling subsets of the cross sectional units is straightforward.

Example 8.25 *Suppose that it is known that*

$$\begin{aligned} \alpha_i &= \bar{\alpha}_1 + v_{i1} \; v_{i1} \sim \mathsf{N}(0, \Sigma_1) \qquad \text{if } \mathrm{i} \leq \mathrm{n}_1 \\ \alpha_i &= \bar{\alpha}_2 + v_{i2} \; v_{i2} \sim \mathsf{N}(0, \Sigma_2) \qquad \text{if } \mathrm{n}_1 > \mathrm{i} > \mathrm{n} \end{aligned} \tag{8.52}$$

Then the four procedures we have described in this subsection can be used to estimate $\bar{\alpha}_1, \bar{\alpha}_2, \Sigma_1, \Sigma_2$ *and* α_i *separately for units in each group.*

Clearly, the assumption that n_1 is known is unrealistic in many applications. Furthermore, standard tests for break points developed in the time series literature are inappropriate for panel data since the ordering of the n units is arbitrary. In chapter 10 we describe how to choose the break point optimally when the ordering of the cross section is unknown.

Exercise 8.39 *Consider a VAR model for unit* i *of the form* $y_{it} = A_i(\ell)y_{it-1} + e_{it}$ *where* $\alpha_i \equiv vec(A_i(\ell)) = \bar{\alpha} + v_i$*. Provide classical and Bayesian estimators of the parameters of the model which combine unit specific and cross sectional information. Is there a reasonable way to check the extent of dynamic heterogeneities?*

8.4.1 What is wrong with two-steps regressions?

There are many situations, both in macroeconomics and in finance, where the parameters of a relationship are assumed to be related to some observable (unit specific) characteristics and researchers employ two-step methods to uncover this relationship.

Example 8.26 *In estimating the cyclicality of government expenditure one may, for example, be interested in knowing if balance budget restrictions matter or not. Therefore, using coefficients estimated from a time series regression as if they were the true ones, a second stage regression on a dummy variable, describing whether a state has a balance budget restrictions or not, is run. Alternatively, in estimating the speed of adjustment of employment to macroeconomic disturbances, one may want to know if labor market institutions account for the empirical differences found. In this case, it is typical to run a regression of the estimated speed of adjustments on cross country indicators of labor market flexibility.*

Is such an approach reasonable? What sort of biases one should expect to find in the second stage estimates? Intuitively, an estimation error is artificially introduced in the second regression and this has important implications. To illustrate why both estimates and standard errors computed from these two-step regressions are incorrect, consider

$$y_{it} = x_{0it}\varrho_i + x_{1it}\alpha_i + e_{it} \tag{8.53}$$

$$\alpha_i = x_{2i}\theta + v_i \tag{8.54}$$

where $i = 1, 2, \ldots, n$ and x_{1it} is a $1 \times m_2$ vector of exogenous and lagged dependent variables, x_{2i} is a $m_2 \times m_3$ vector of time invariant unit specific characteristics and x_{0it} is a $1 \times$

m_1 vector of unit specific intercepts (possibly depending on t). Finally, θ is a $m_3 \times 1$ vector of parameters. We assume that $E(x_{1it}e_{it}) = E(x_{2i}v_i) = 0$ that $e_{it} \sim \mathsf{N}(0, \sigma_i^2)$; that $E(e_{it}, e_{i'\tau}) = 0 \ \forall t \neq \tau$ and $i \neq i'$; and $v_i \sim \mathsf{N}(0, \Sigma_v)$. Stacking the observations for each i and using (8.54) into (8.53) we have $y_i = x_{0i}\varrho_i + X_i\theta + \epsilon_i$ where $X_i = x_{1i}x_{2i}$ is a $T \times m_3$ matrix, and $\epsilon_i = x_{1i}v_i + e_i$ so that $\text{var}(\epsilon_i) = x_{1i}\Sigma_v x_{1i}' + \sigma_i^2 I \equiv \Sigma_{\epsilon_i}$.

Exercise 8.40 *Show that given Σ_{ϵ_i} and θ the ML estimator of ϱ_i is $\varrho_{i,ML} = (x_{oi}'\Sigma_v^{-1}x_{0i})^{-1}(x_{oi}'\Sigma_v^{-1}(y_i - x_i\theta)$ and that conditional on Σ_{ϵ_i} $\theta_{ML} = (\sum_i X_i\Omega_i X_i)^{-1}(\sum_i X_i\Omega_i y_i)$ where $\Omega_i = (\Sigma_{\epsilon_i}^{-1} - \Sigma_{\epsilon_i}^{-1}x_{0i}(x_{0i}\Sigma_{\epsilon_i}^{-1}x_{0i})^{-1}x_{0i}'\Sigma_{\epsilon_i}^{-1})$*

Using the same logic of exercise 8.37, we can write $\theta_{ML} = (\sum_i x_{2i}'\tilde{\Omega}_i^{-1}x_{2i})^{-1})(\sum_i x_{2i}'\tilde{\Omega}_i^{-1}\hat{\alpha}_i)$ where $\tilde{\Omega} = (x_{1i}'x_{1i})^{-1}\Omega_i$. Therefore the maximum likelihood estimate of θ, which corresponds to the GLS estimate of the transformed model, is a weighted average of the first stage estimates $\hat{\alpha}_i$ with weights which depend on Ω.

It is easy to see that second stage estimates of θ are $\theta_{2step} = (\sum_i x_{2i}'\Sigma_v^{-1}x_{2i})(\sum_i x_{2i}'\Sigma_v^{-1}\hat{\alpha}_i)$. Therefore θ_{2step} incorrectly measures the effect of x_{2i} on α_i for two reasons. First, suppose that $x_{i0t} = 0, \ \forall t$. Then θ_{2step} neglects the fact that α_i are estimated (i.e. it neglects the term $\sigma^2(x_{1i}'x_{1i})^{-1}$). Moreover, the weights used in θ_{2step} are homoschedastic while those in θ_{ML} depend on unit specific regressors x_{1i}. Second, if $x_{i0t} \neq 0$, there are additional terms in Ω_i which θ_{2step} neglects. It is difficult to predict what the combined effect of these two errors would be. In general, treating estimates as if they were the true ones will make θ_{2step} artificially significant and, in particular situations, may also bias the sign of the relationship.

Given that ML estimates are easy to compute and are feasible once estimates of σ_i^2 and Σ_v are plugged in the formulas, there is no reason to prefer two-steps estimators. The mismeasurement caused by a two-steps approach can be important as it is shown next.

Example 8.27 *We use US state data to estimate whether the cyclicality of government expenditure share in states with strict (ex-post) balance budget restrictions is different from that of states with weak (ex-ante) balance budget restrictions. We use annual data for 48 states (and 13 have only weak restrictions) from 1969 to 1995 and compute two regressions: one with a two step model, i.e. $\ln \frac{G_{it}}{GDP_{it}} = \varrho_i + \alpha_{1i} \ln \frac{G_{it-1}}{GDP_{it-1}} + \alpha_{2i}\Delta \ln GDP_{it} + e_{it}$ and $\alpha_{2i,ols} = BB\theta_1 + (1 - BB)\theta_2 + v_i$ where BB_i is a dummy variable taking the value of 1 if strict restrictions are present and zero otherwise; and one with a one-step model i.e. $(1-\alpha_{1i}\ell) \ln \frac{G_{it}}{GDP_{it}} = \varrho_i + \theta_1(BB\Delta \ln GDP_{it}) + \theta_2((1-BB)\Delta \log GDP_{it}) + \epsilon_{it}$. The estimates of the coefficients of the two regressions are the same $\theta_1 = 0.81$, $\theta_2 = 0.54$, suggesting a larger cyclicality for states without balance budget constraints. However while with the two-steps regression the standard error of the estimates are 0.09 and 0.07, they are 1.58 and 1.87 with the one-step regression. Hence, an asymptotic t-test for equality of the effect has a p-value of 0.08 with the two-step regression and of 0.83 with the one-step regression.*

8.5 Is Money superneutral?

To illustrate some of the issues discussed in this chapter we study the effects of money on output in the long run using the cross section of G-7 countries. The majority of monetary dynamic general equilibrium models have built in some form of money neutrality so that, in the long run, real variables are insulated from nominal ones. However, in some cash-in-advance models, variations of the growth rate of money may have real effects, even in the long run, because they alter the marginal rate of substitution between consumption and leisure and induce agents to work less in the steady state.

Example 8.28 *Consider the cash-in advance model described in example 1.4 of chapter 2. Suppose we select as instantaneous utility function* $u(c_{1t}, c_{2t}, N_t) = \vartheta_c \log c_{1t} + (1 - \vartheta_c) \log c_{2t} - \vartheta_N N_t$, *where* c_{1t} *are cash goods and* c_{2t} *are credit goods. Letting the growth rate of money* ΔM *follow an AR(1) process with mean* $\bar{M}$, *persistence* ρ_M *and standard deviation* σ_M, *log linearizing the conditions around the steady state and setting* $\beta = 0.989, \delta = 0.019, \eta = 0.6, \vartheta_N = 2.53, \vartheta_c = 0.4, \bar{M} = 0.015, \rho_\zeta = 0.95, \rho_M = 0.45, \sigma_\zeta^2 = 0.07, \sigma_M^2 = 0.0089$, *the decision rule for hours is* $\ln N = 0.25 + 1.51\zeta_t - 0.05\Delta M_t - 0.45 \ln K_t$. *Hence, ceteris paribus, increases in the growth rate of money have depressing short run effects on hours worked and therefore, via the production function, real activity. However, money growth disturbances also affect the steady state of the economy. In fact,* $\rho_M = 0$, *the compensating variations in consumption needed to bring agents back to the optimum, are 0.520 when* $\bar{M} = 0.10$ *and 0.972 when* $\bar{M} = 0.20$.

The literature has thoroughly discussed the problems one may encounter when using variations in the growth rate of money to proxy for monetary policy actions (see e.g. Gordon and Leeper (1994)). First, there are variations which may represent responses to the state of the economy. Second, even when innovations in growth rates are considered, they may capture demand variations, as opposed to supply changes. With these caveats in mind we examine whether innovations in money growth have long run effects on output growth in the G-7 using three different estimators. In one case we average the responses of output growth to money growth shocks across countries; in the second, we compute one average response pooling the data across countries; in the third, we aggregate cross sectional data for each t and compute the response of output growth to money growth shocks. Our interest here is multiple. First, we would like to see if different estimators tell us different stories about the superneutrality of money. Second, we want to relate differences, when they exist, to the properties of estimators. Third, we want to see if the evidence is consistent with the prediction of the model of example 8.28.

The data covers the sample 1980:1-2000:4, is demeaned to eliminate fixed effects, and for each specification we run a bivariate VAR(5). We make the somewhat heroic assumption that no variable, other than output or money, is helpful in understanding the relationship between these two variables. Given the approximate diagonality of the covariance matrix of reduced form shocks for all countries, the identifying restrictions imposed to separate money growth shocks from output growth shocks give the same result.

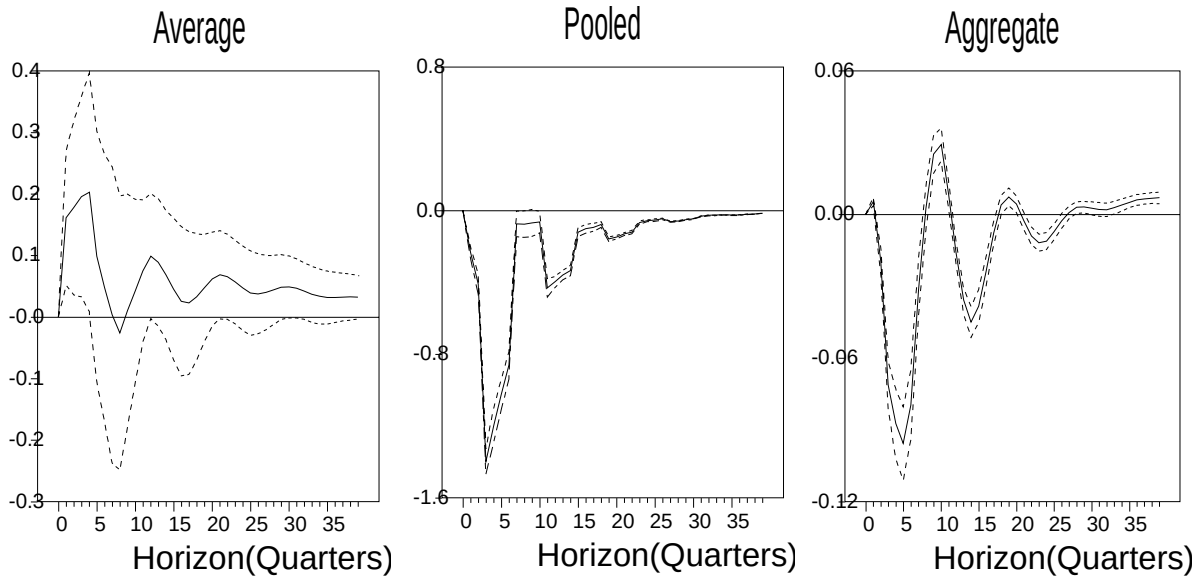


Figure 8.3: Output growth responses

Figure 8.3 indicates that different estimators produce different responses. For example, when the pooled and the aggregate estimators are used (both of which are inconsistent when dynamic heterogeneity is present). In both cases we observe negative short run output growth responses and a jagged pattern in the medium run. With the average estimator (which is consistent with dynamic heterogeneity) responses are consistently positive, albeit insignificant after about a year. We checked whether dynamic heterogeneity is important by testing the orthogonality conditions implied by the pooled and the average estimators. The smaller statistic equals to 58.23, so that the null of homogeneity is soundly rejected when compared with a $\chi^2(10)$. This is unsurprising: it is well known that the money growth path of Italy and the UK had very different properties than the path of, say, Germany or Japan in the 1980s. What is remarkable is that with pooled or aggregate estimates one may be led to accept some of the predictions of the simple cash-in-advance model of example 8.28, while the opposite would occur when the average estimator is used.

We examine long run superneutrality in two ways. First, we check if output growth responses at the 10 year horizon are statistically significant. Second, we examine whether the contribution of money growth shocks to the variance of output growth at the same horizon is economically significant. Differences across estimators also emerge with these statistics. The average and the pooled estimator produce insignificant responses while responses obtained with the aggregate estimator are statistically significant. The economic differences are however small. In fact, in the latter case, a 68% band for the contribution of money growth shocks to output growth variability at the 10 year horizon is [0.02,0.14], which covers almost entirely the band obtained with the other two estimators.

Next, we examine the contribution of cross sectional information when measuring the responses of output growth to money growth shocks in Japan. Figure 8.4 presents point estimates of the responses obtained using (a) only local information, (b) a Stein estimator, (c) a random coefficient estimator, where the a-priori mean of all the responses is zero and the prior variances are 0.05. The response obtained using local information is hump shaped and positive throughout the range, converging to zero rather slowly. This pattern is maintained with the random coefficient estimator but the peak response is reduced and the response is smoother. The response obtained with the Stein estimator is oscillatory, reflecting the jagged pattern of the pooled estimator (see figure 8.3). Note also that while in the first two cases responses are significant up to the 10 years horizon, in the latter they are statistically significant only from 7 to 10 quarters after the shock.

To conclude, money appears to be superneutral in the long run and economic deviations from the null hypothesis are small. The short and the medium run response of output growth to money growth shocks depends on the estimation technique. Since with dynamic heterogeneity two of the estimators are inconsistent, we conclude that, on average in the G-7, money growth has a positive short run effect on output growth lasting about one year.

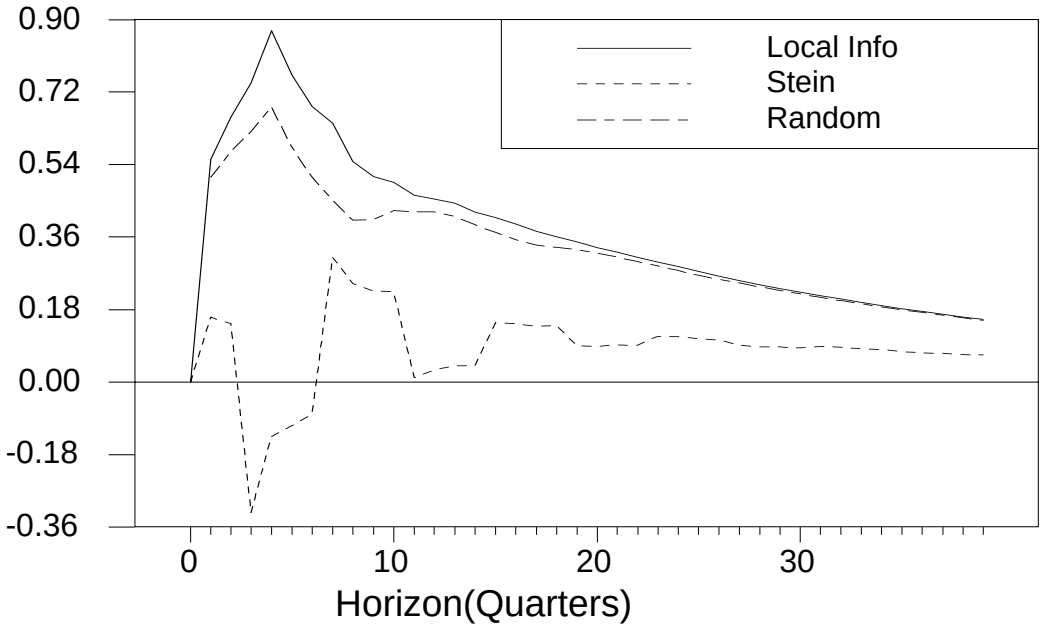


Figure 8.4: Alternative Estimators of Output Responses in Japan

Chapter 9: Introduction to Bayesian Methods

Bayesian analysis of statistical and economic models differs substantially from the classical (frequentist) one. In classical analysis the probability of an event is the limit of the relative frequency of that event. Furthermore, the parameters of a model are treated as fixed, unknown quantities. In this framework, unbiased estimators are important because the average value of the sample estimator converges to the true value via some Law of Large Numbers. Also, minimum variance estimators are preferable because they yield values closer to the true parameter. Finally, estimators and tests are evaluated in repeated samples since this insures that they give correct results with high probability.

Bayesian analysis takes a different point of view on all these issues. Probabilities measure the degree of beliefs that a researcher has in an event. Parameters are random variables with a probability distribution. Properties of estimators and tests in repeated samples are uninteresting since beliefs are not necessarily related to the relative frequency of an event in large number of hypothetical experiments. Finally, estimators are chosen to minimize expected loss function (with expectations taken with respect to the posterior distribution), conditional on the data.

Despite these philosophical and semantic differences, it turns out that the two procedures are equivalent in large samples. That is, under regularity conditions, the posterior mode converges to some " true" parameter value as $T \to \infty$; furthermore, the posterior distribution converges to a normal with mean equal to the "true" parameter and variance which is proportional to Fisher's information matrix.

This chapter describes the basics of Bayesian analysis. These tools are fundamental for the study of interesting macroeconomic problems encountered in the next two chapters. The building block of the analysis is the Bayes theorem. Section 1 presents examples of how Bayes theorem can be used to construct posterior distributions and to recursively update posterior information. Crucial for the analysis is the specification of a prior distribution for the parameters. We describe ways of selecting such a prior, distinguishing between subjective and objective approaches; contrast informative and non-informative priors and present conjugate priors, which play an important role in many applied problems.

The second section deals with decision theory and the third with inference. We describe how to obtain point and interval posterior estimates, discuss asymptotic properties

of Bayesian estimators; how to compare hypotheses/models and to construct distributions of forecasts. The fourth section deals with hierarchical models. Such models feature either a two-stage prior or a latent variable structure, which naturally lends itself to a prior interpretation. Since models of this type are common in applied work, we describe in details the steps needed to construct posterior distributions in these setups. We will also discuss Empirical Bayes methods here. These methods use plug-in estimates of the parameters of a second stage hierarchy to construct posterior estimates of the first stage parameters; they are useful for complex problems where the construction of the joint posterior of first and second stage parameters is demanding, and have a number of applications, in particular to VARs. Section 5 deals with posterior simulators. When the form of the posterior distribution is unknown, one can conduct posterior analysis drawing sequences from a distribution which approximates the posterior. We discuss normal approximations; more sophisticated acceptance/importance sampling approximations and recent Markov Chain Monte Carlo methods. These latter methods are powerful and will be extensively used in the multi-parameters, hierarchical, non-linear, state space and latent variable models considered in chapters 10 and 11.

Section 6 briefly deals with robustness. Whenever samples are short, one is interested in knowing how important is the prior in determining the shape of the posterior. We describe a simple approach to assess the importance of a prior specification and to provide readers and clients with ways to rebalance posterior information in a way that suits their purposes. Section 7 applies some of the tools described in the chapter to the problem of measuring returns to scale in the Spanish production function.

9.1 Preliminaries

Throughout this chapter, we assume that the parameters of interest α lie in a compact set A. Prior information is summarized by a density $g(\alpha)$. Sample information is represented by a density $f(y|\alpha) \equiv \mathcal{L}(\alpha|y)$, which can be interpreted as the likelihood of α once the data y is observed. $\tilde{\alpha}$ represents a posterior estimator of α and α_{ols} its sample estimator.

9.1.1 Bayes Theorem

Bayes Theorem allows to compute the posterior of α from the prior and the likelihood:

$$g(\alpha|y) = \frac{f(y|\alpha)g(\alpha)}{f(y)} \propto f(y|\alpha)g(\alpha) = \mathcal{L}(\alpha|y)g(\alpha) \equiv \breve{g}(\alpha|y)$$

where $f(y) = \int f(y|\alpha)g(\alpha)d\alpha$ is the predictive density; $g(\alpha|y)$ is the posterior density of α, and $\breve{g}(\alpha|y)$ is the posterior kernel. By construction, $g(\alpha|y) = \frac{\breve{g}(\alpha|y)}{\int \breve{g}(\alpha|y)d\alpha}$.

Example 9.1 *Suppose a Central Bank has to decide on an interest rate policy; it can choose to increase the rate, to decrease it or leave it unchanged. In each of the three cases a recession may occur or not. Let i_1 indicate an interest rate increase, i_2 an unchanged*

interest rate and i_3 an interest rates decrease. Let Re be a recession event and NRe a non-recession event. Suppose that $f(Re|i_1) = 0.5$; $f(Re|i_2) = 0.4$; $f(Re|i_3) = 0.3$, and that all interest rate policies are equally probable a-priori. Then, the probability that the interest rate will decrease, given that a recession is observed, is $g(i_3|Re) = (0.3 * 0.33)/(0.5 * 0.33 + 0.4 * 0.33 + 0.3 * 0.33) = 0.68$ and the probability that the interest rate will increase, given that no-recession occurred, is $g(i_1|NRe) = (0.5 * 0.33)/(0.5 * 0.33 + 0.6 * 0.33 + 0.7 * 0.33) = 0.75$.

The next example uses the Bayes theorem to recursively update prior information.

Example 9.2 *Suppose you are betting on the winner of the soccer World Cup. Your team is Brazil. Let α_1 represent the event that it will win the championship and α_2 the event it will loose it. Suppose Brazil meets Spain in a robin round. Your prior is $g(\alpha_1) = 0.6$, $g(\alpha_2) = 0.4$. Let $y_1 = 1$ if Brazil wins the game and $y_1 = 0$ otherwise. Suppose that $f(y_1 = 1|\alpha_1) = 0.8$, $f(y_1 = 1|\alpha_2) = 0.3$. Then $f(y_1 = 1) = 0.8 * 0.6 + 0.3 * 0.4 = 0.6$; where $f(y_1 = 1)$ is the proportions of wins one can anticipate. Then*

$$g(\alpha_1|y_1 = 1) = \frac{f(y_1 = 1|\alpha_1)g(\alpha_1)}{f(y_1 = 1)} = \frac{0.8 * 0.6}{0.6} = 0.8 \tag{9.1}$$

$$g(\alpha_2|y_1 = 1) = \frac{f(y_1 = 1|\alpha_2)g(\alpha_2)}{f(y_1 = 1)} = \frac{0.3 * 0.4}{0.6} = 0.2 \tag{9.2}$$

Hence, having beaten Spain, the probability Brazil will win the championship increases from 0.6 to 0.8. Suppose that the next opponent is Cameroon. Let $y_2 = 1$ if Brazil wins this game and $y_2 = 0$ otherwise. Let $g(\alpha_1) = g(\alpha_1|y_1 = 1) = 0.8$ and $g(\alpha_2) = g(\alpha_2|y_1 = 1) = 0.2$, that is, the prior at this new stage is the posterior of the previous stage. Let again $f(y_2 = 1|\alpha_1) = 0.8$, $f(y_2 = 1|\alpha_2) = 0.3$. Then $f(y_2 = 1|y_1 = 1) = 0.8 * 0.8 + 0.3 * 0.2 = 0.7$ and

$$g(\alpha_1|y_1 = 1, y_2 = 1) = \frac{f(y_2 = 1|\alpha_1, y_1 = 1)g(\alpha_1|y_1 = 1)}{f(y_2 = 1|y_1 = 1)} = \frac{0.64}{0.7} = 0.91 \tag{9.3}$$

$$g(\alpha_2|y_1 = 1, y_2 = 1) = \frac{f(y_2 = 1|\alpha_2, y_1 = 1)g(\alpha_2|y_1 = 1)}{f(y_2 = 1|y_1 = 1)} = \frac{0.06}{0.7} = 0.09 \tag{9.4}$$

Having beaten Cameroon, Brazil has now 0.91 probability of winning the championship.

Exercise 9.1 *Consider the case of a chip which goes into a computer: it can be either proper or faulty. Still, in each of the two cases, the computer may or may not work. Let α_1 be the event that the chip is proper, α_2 the event that the chip is faulty and let y be the event that the computer is working. Suppose it is known that $f(y|\alpha_1) = 0.995$ (that is, the probability a computer works when the chip is proper) and $f(y|\alpha_2) = 0.005$. Previous records show that $g(\alpha_1) = 0.997$ and $g(\alpha_2) = 0.003$. Calculate the probability that the chip is proper, given that the computer under consideration is working, i.e. compute $g(\alpha_1|y)$.*

Exercise 9.2 *Consider n independent draws of a two-point event (e.g. high vs. low inflation or high vs. low growth rate of M3). Let α be the probability that a high event*

occurs. If n_1 episodes of the high event are observed, the likelihood function is $f(n_1|\alpha, n) = \frac{n!}{n_1!(n-n_1)!}\alpha^{n_1}(1-\alpha)^{n-n_1}$. Suppose that $g(\alpha) = (\alpha(1-\alpha))^{-1}$ for $0 \leq \alpha \leq 1$). Show the form of the posterior density for α. What is the mean of the posterior distribution? What is the mode? Calculate $g(\alpha|n_1 = 5, n = 20)$.

Exercise 9.3 *Consider two types of workers, high skilled (Hs) and low skilled (Ls) and let their employment status be Em (employed) or Un (unemployed). Suppose that historically $P(Em|Hs, Em) = 0.85, P(Em|Ls, Em) = 0.6, P(Un|Hs, Un) = 0.3, P(Un|Ls, Un) = 0.7$ so that, e.g., the probability that a low skilled (high skilled) unemployed will find a job this period is 0.3 (0.7). Suppose that a priori it is known that job flows are such that $P(Un) = 0.4$ and $P(Em) = 0.6$. Calculate the posterior probability that a low skilled unemployed worker will still be unemployed two periods from now. How would this probability change if a training program alters $P(Em|Ls, Un)$ from 0.3 to 0.4?*

There are many situations when the vector α contains nuisance parameters (parameters which are of little importance for the goal of the investigation). Posterior distributions for the objects of interest when nuisance parameters are present, can be easily computed. Suppose $\alpha = [\alpha_1, \alpha_2]$ and suppose we are interested only in α_1. The joint posterior is $g(\alpha_1, \alpha_2|y) \propto f(y|\alpha_1, \alpha_2)g(\alpha_1, \alpha_2)$ and $g(\alpha_1|y) = \int g(\alpha_1, \alpha_2|y)d\alpha_2 = \int g(\alpha_1|\alpha_2, y)g(\alpha_2|y)d\alpha_2$, that is, the marginal posterior of α_1 weights the conditional posterior of α_1 with the marginal posterior of α_2. When the dimension of α_2 is large, integrating α_2 out of the joint posterior is difficult. In this case, one could use Monte Carlo methods to obtain an iid sequence from the posterior. Suppose $g(\alpha_2|y)$, and $g(\alpha_1|y, \alpha_2)$ are available. Then the following can be used:

Algorithm 9.1

1) *Draw α_2^l from $g(\alpha_2|y)$. For each α_2^l, draw $\alpha_1^{l'}$ from $g(\alpha_1|y, \alpha_2^l)$, $l' = 1, \ldots, L'$.*

2) *Average $\alpha_1^{l'}$ over draws, i.e compute $\alpha_1 = \frac{1}{L'}\sum_{l'} \alpha_i^{l'}$.*

3) *Repeat 1)- 2) L times.*

The sample $(\alpha_1^1, \ldots, \alpha_1^L)$ is then a sequence from the marginal posterior $g(\alpha_1|y)$.

9.1.2 Prior Selection

The Bayes theorem requires the specification of a prior density $g(\alpha)$. At one extreme, $g(\alpha)$ may represent the subjective beliefs a researcher has in the occurrence of an event (e.g., the probability that there is a defective CD in a lot of 1000). At the other, it may represent an objective evaluation: it may reflect recorded information (e.g. how many times has a lightning storm occurred in Rome on August 15 in the last 100 years), or the outcomes of previous experiments. Half way in between are priors displaying subjective general features (e.g. the form of the distributions) and objective details (e.g. the moments). Priors can also be distinguished on the basis of their informational content. In this case, we classify a prior as informative or non-informative.

Subjective Priors

Subjective informative priors can be constructed in a number of ways. For example, one can split the support of α into intervals, attribute a probability to each interval and connect piecewise the intervals (histogram approach). Alternatively, one subjectively computes the "likelihood" of various $\alpha \in A$ and connect the likelihood points (likelihood approach). Finally, one can take moments (or fractiles) of $g(\alpha)$ as given and choose the functional form for $g(\alpha)$ that best described by these moments (or fractiles) (functional form approach).

Non-informative priors are typically selected when information is scarce or when a researcher wants to minimize the influence of the prior on the posterior. When subjective non-informative priors are chosen, it is typical to require them to be informationally invariant, i.e. if $g(\alpha) = \kappa$ is non-informative, then $g(\alpha) = \varrho\kappa$ should also be non-informative, where ϱ is a constant.

The literature has developed what are called *reference non-informative priors*, i.e priors which are invariant either in their location, in their scale or in both. In the next example we describe how to obtain a location invariant prior in a general case.

Example 9.3 *Suppose* R_1 *and* R_2 *are subsets of* R^m*, that the density of* y *is of the form* $f(y-\alpha)$ *where* $\alpha \in \mathsf{R}_2$ *is a location parameter and* $y \in \mathsf{R}_1$*. For example, a normal distribution with mean* α *and known variance* σ^2 *is a location distribution. To derive an invariant non-informative prior suppose that instead of observing* y*, we observe* $y_1 = y + \varrho$*,* $\varrho \in \mathsf{R}^m$*. Letting* $\alpha_1 = \alpha + \varrho$*, the density of* y_1 *is of the form* $f(y_1 - \alpha_1)$*. Hence, since the densities of* (y, α) *and* (y_1, α_1) *are identical in structure they must have the same non-informative prior; that is, we want*

$$g(\alpha \in \mathsf{R}_2) = g(\alpha_1 \in \mathsf{R}_2) \tag{9.5}$$

for all $\mathsf{R}_2 \in \mathsf{R}^m$*. Since* $g(\alpha_1 \in \mathsf{R}_2) = g(\alpha + \varrho \in \mathsf{R}_2) = g(\alpha \in \mathsf{R}_2 - \varrho)$*, where* $\mathsf{R}_2 - \varrho = \{z - \varrho : z \in \mathsf{R}_2\}$*, we have that* $g(\alpha \in \mathsf{R}_2) = g(\alpha \in \mathsf{R}_2 - \varrho)$*. Since* ϱ *is arbitrary, a prior satisfying this equality is a location invariant prior. Integrating the above expression we have*

$$\int_{\mathsf{R}_2} g(\alpha)d\alpha = \int_{\mathsf{R}_2-\varrho} g(\alpha)d\alpha = \int_{\mathsf{R}_2} g(\alpha-\varrho)d\alpha \tag{9.6}$$

which is true if and only if $g(\alpha) = g(\alpha - \varrho)$*,* $\forall\alpha$*. Setting* $\alpha = \varrho$ *we have* $g(\alpha) = g(0)$*,* $\forall\varrho$*, i.e.* g *must be a constant. For convenience, it is typical to choose* $g(\alpha) = 1$*.*

Exercise 9.4 *Consider a scale density of the form* $\sigma^{-1}f(\frac{y}{\sigma})$ *where* $\sigma > 0$ *(e.g.* $y \sim \mathsf{N}(0, \sigma^2)$ *is a scale density). Show that an invariant non-informative prior is* $g(\sigma) = \sigma^{-1}$*. (Hint: repeat the steps of example 9.3, assuming you observe* $y_1 = \varrho y$*,* $\varrho \in \mathsf{R}^m$*).*

Although reference priors are often taken off-the shelf, they are not automatic since they depend on the parametrization of the model. For example, while $g(\sigma) = \sigma^{-2}$ is non-informative for σ^2, $g(\sigma_1) = 1$ is non-informative for $\sigma_1 = log(\sigma^2)$.

Jeffrey (1966) proposed a method to generate non-informative priors based on Fisher information matrix, i.e. the matrix of expected second order derivatives of the density of the data with respect to the parameters. The idea is simple: let $g(\alpha)$ be given and let $\alpha^0 = h(\alpha)$, where h is continuous and differentiable function. Then, the prior for α^0 is $g(\alpha^0) = g(\alpha)|\frac{\partial h(\alpha)}{\partial \alpha}|$. Jeffrey's principle states that the $g(\alpha^0)$ obtained through this transformation must be the same as the one obtained using $g(\alpha^0) = \frac{g(\alpha^0, y)}{f(y|\alpha^0)}$. If $g(\alpha) \propto [I(\alpha)]^{0.5}$, where $I(\alpha) = -E[\frac{\partial^2 \ln f(y|\alpha)}{\partial \alpha^2}|\alpha]$, then the same prior for α^0 obtains with the two approaches.

Example 9.4 *Let α represents the probability that the growth rate of output y_t is above average and $1-\alpha$ the probability that it is below average at each t. Then $f(y_t|\alpha) = \alpha^{y_t}(1-\alpha)^{1-y_t}$ and $-E[\frac{\partial^2 \ln f(y_t|\alpha)}{\partial \alpha \partial \alpha'}] \approx E([\frac{\partial \ln f(y_t|\alpha)}{\partial \alpha}]^2) = E[(\alpha^{-1}y_t + (1-\alpha)^{-1}(y_t - 1))^2] = \alpha^{-1}(1-\alpha)^{-1}$. Hence, Jeffrey's prior for α is $g(\alpha) = \alpha^{-0.5}(1-\alpha)^{-0.5}$.*

Example 9.5 *Consider the linear regression model $y = x\alpha + e$, where α is a scalar and $e \sim \mathsf{N}(0,1)$. Then $f(y|\alpha, x) = \frac{1}{\sqrt{2\pi}} exp\{-0.5(y - x\alpha)^2\}$ and $\frac{\partial^2 \log f(y|\alpha)}{\partial \alpha^2} = -x^2$ so $g(\alpha) \propto E(x)$ is an invariant prior for α. Note that this prior is data based.*

Exercise 9.5 *Consider a location-scale density of the form $\sigma^{-1} f(\frac{y-\alpha}{\sigma})$ where (α, σ) are unknown and assume that $f(\frac{y-\alpha}{\sigma}) \propto exp\{-0.5\sigma^{-2}(y-\alpha)^2\}$. Show that Fisher information is $I(\alpha, \sigma^2) =$ diag $(\frac{1}{\sigma^2}, \frac{3}{\sigma^2})$, so that a non-informative prior for (α, σ^2) is $g(\alpha, \sigma^2) \propto \sigma^{-2}$.*

It is straightforward to extend Jeffrey's formulation to the multivariate case. In fact, if $\alpha = (\alpha_1, \ldots, \alpha_k)$, then $g(\alpha) = \{\det[I(\alpha)]\}^{0.5}$ where $I_{i,j}(\alpha) = -E_\alpha[\frac{\partial^2 \log f(y|\alpha)}{\partial \alpha_i \alpha_j}|\alpha]$.

At times the terms non-informative and improper are used interchangeably to indicate priors with diffuse information but the two concepts are distinct. Since a prior is improper if $g(\alpha) \geq 0$, $\forall \alpha \in A$ but $\int_A g(\alpha)d\alpha$ is divergent, it is easy to construct examples where the two terms are not interchangeable. That is, an improper prior may be non-informative under one parametrization and informative under another.

Example 9.6 *Suppose α is scalar, A the real line and $g(\alpha) = 1$, $\forall \alpha \in A$. This prior is non-informative and improper. Consider the reparameterization $\alpha_1 = \frac{exp(\alpha)}{1+exp(\alpha)}$. Then $\alpha_1 \in (0,1)$, and $g(\alpha_1) \propto (\alpha_1)^{-1}(1-\alpha_1)^{-1}$, which is informative (it is heavily concentrated around 0 and 1) but still improper.*

Objective priors

In formulating objective priors the predictive density of the data plays a crucial role. Such a density measures the likelihood of y based on sample information and is the normalizing constant in the Bayes theorem, i.e.:

$$f(y) = \int_A f(y|\alpha)g(\alpha)d\alpha \equiv \mathcal{L}(y|g) \tag{9.7}$$

Example 9.7 *Suppose that y represents the number of papers a researcher publishes in the American Economic Review (AER) in the lifetime and let y be normally distributed around an unobservable ability variable α. Suppose that abilities in the population vary according to a normal distribution with mean $\bar{\alpha}$ and variance σ_a^2. Then $f(y)$ represents the actual distribution of AER articles of that researcher. Any idea what is the median of this distribution in the cross section of economists? Zero!*

Important insights can be gained with a closer look at (9.7). Since $f(y|\alpha)$ is fixed, $\mathcal{L}(y|g)$ reflects the plausibility of g in the data. Therefore, if g_1 and g_2 are two prior distributions, $\mathcal{L}(y|g_1) > \mathcal{L}(y|g_2)$ implies that the support for g_1 in the data is larger than the one for g_2. Taken one step further this idea implies that we can estimate the "best" g using $\mathcal{L}(y|g)$. In fact, let $g(\alpha) \equiv g(\alpha|\theta)$ be a function of some hyperparameters θ. Then $\mathcal{L}(y|g) \equiv \mathcal{L}(y|\theta)$ and θ_{ML}, typically called maximum likelihood type II (ML-II) estimator, is the θ which maximizes $\mathcal{L}(y|\theta)$ and $g(\alpha|\theta_{ML})$ is a ML-II based prior.

Example 9.8 *Let $y|\alpha \sim \mathsf{N}(\alpha, \sigma_y^2)$, $\alpha \sim \mathsf{N}(\bar{\alpha}, \bar{\sigma}_a^2)$, σ_y^2 known. Then $\mathcal{L}(y|g) \sim \mathsf{N}(\bar{\alpha}, \sigma^2 = \sigma_y^2 + \bar{\sigma}_a^2)$. If T observations are available, $\mathcal{L}(y_1, \dots y_T|g)$ can be written as*

$$\mathcal{L}(y_1, \dots y_T|g) = [2\pi(\sigma^2)]^{-0.5T} \exp\{-0.5\frac{Ts^2}{\sigma^2}\} \exp\{-0.5\frac{T(\bar{y}-\bar{\alpha})^2}{\sigma^2}\} \tag{9.8}$$

where $\bar{y} = \frac{1}{T} \sum_t y_t$, $s^2 = \frac{1}{T} \sum_t (y_t - \bar{y})^2$. Maximizing (9.8) with respect to $\bar{\alpha}$ yields $\bar{\alpha}_{ML} = \bar{y}$. Substituting this into (9.8) we obtain

$$\mathcal{L}(y_1, \dots y_T|\bar{\alpha}_{ML}, g) = [2\pi(\sigma_y^2 + \bar{\sigma}_a^2)]^{-0.5T} \exp\{-0.5\frac{Ts^2}{\sigma_y^2 + \bar{\sigma}_a^2}\} \tag{9.9}$$

Maximizing (9.9) with respect to $\bar{\sigma}_a^2$ we have $\bar{\sigma}_{aML}^2 = s^2 - \sigma_y^2$ when $s^2 \geq \sigma_y^2$ and $\bar{\sigma}_{aML}^2 = 0$ otherwise. Hence a ML-II prior for α is normal with mean $\bar{\alpha}_{ML}$ and variance $\bar{\sigma}_{aML}^2$.

Conjugate Priors

Conjugate priors are convenient because they allow the analytical computation of the posterior distribution of the unknowns in simple models.

Definition 9.1 *Let F be a class of sampling distributions and G a class of prior distributions. G is conjugate for F if $g(\alpha|y) \in \mathsf{G}$, for all $f(y|\alpha) \in \mathsf{F}$ and $g(\alpha) \in \mathsf{G}$.*

A prior is conjugate if the beliefs it represents can be described by a density which has the same form as the actual data. Since the posterior is proportional to the prior times the likelihood, it will have the same form as the data and the prior.

Example 9.9 *Let $y \sim \mathsf{N}(\alpha, \sigma^2)$, σ^2 known. Since $f(y|\alpha) = \frac{1}{\sigma\sqrt{2\pi}} exp\{-\frac{1}{2\sigma^2}(y-\alpha)^2\}$ a conjugate prior for α must be quadratic in the exponent, i.e: $g(\alpha) \propto exp\{-A_0\alpha^2 - A_1\alpha - A_2\}$,*

where A_0, A_1, A_2 are constants. Set $g(\alpha) = \frac{1}{\sigma_\alpha\sqrt{2\pi}}\exp\{-\frac{1}{2\sigma_\alpha^2}(\alpha-\bar{\alpha})^2\}$, $\bar{\alpha}, \bar{\sigma}_\alpha$ known. Then $f(y,\alpha) = \frac{1}{2\pi\sigma\sigma_\alpha}exp\{-0.5(\frac{(\alpha-\bar{\alpha})^2}{\bar{\sigma}_\alpha^2} + \frac{(y-\alpha)^2}{\sigma^2})\}$. Note that

$$\begin{aligned}\frac{(\alpha-\bar{\alpha})^2}{\bar{\sigma}_\alpha^2} + \frac{(y-\alpha)^2}{\sigma^2} &= \alpha^2(\frac{1}{\bar{\sigma}_\alpha^2}+\frac{1}{\sigma^2}) - 2\alpha(\frac{y}{\sigma^2}+\frac{\bar{\alpha}}{\bar{\sigma}_\alpha^2}) + (\frac{y^2}{\sigma^2}+\frac{\bar{\alpha}^2}{\bar{\sigma}_\alpha^2}) \\ &= (\bar{\sigma}_\alpha^{-2}+\sigma^{-2})(\alpha^2 - \frac{2\alpha}{\bar{\sigma}_\alpha^{-2}+\sigma^{-2}}(\frac{y}{\sigma^2}+\frac{\bar{\alpha}}{\bar{\sigma}_\alpha^2})) + (\frac{y^2}{\sigma^2}+\frac{\bar{\alpha}^2}{\bar{\sigma}_\alpha^2})\end{aligned}$$

Since $(\alpha^2 - \frac{2\alpha}{\bar{\sigma}_\alpha^{-2}+\sigma^{-2}}(\frac{y}{\sigma^2}+\frac{\bar{\alpha}}{\bar{\sigma}_\alpha^2}) = (\alpha - \frac{1}{\bar{\sigma}_\alpha^{-2}+\sigma^{-2}}(\frac{y}{\sigma^2}+\frac{\bar{\alpha}}{\bar{\sigma}_\alpha^2}))^2 - \frac{1}{\bar{\sigma}_\alpha^{-2}+\sigma^{-2}}(\frac{y^2}{\sigma^2}+\frac{\bar{\alpha}^2}{\bar{\sigma}_\alpha^2})^2$, we have that

$$\frac{(\alpha-\bar{\alpha})^2}{\bar{\sigma}_\alpha^2} + \frac{(y-\alpha)^2}{\sigma^2} = (\bar{\sigma}_\alpha^{-2}+\sigma^{-2})(\alpha - \frac{1}{\bar{\sigma}_\alpha^{-2}+\sigma^{-2}}(\frac{y}{\sigma^2}+\frac{\bar{\alpha}}{\bar{\sigma}_\alpha^2}))^2 + \frac{(y-\bar{\alpha})^2}{(\sigma^2+\bar{\sigma}_\alpha^2)} \quad (9.10)$$

so that $f(y,\alpha) = \frac{1}{\sqrt{2\pi}\sigma\bar{\sigma}_\alpha}exp\{-0.5(\bar{\sigma}_\alpha^{-2}+\sigma^{-2})[\alpha - \frac{1}{\bar{\sigma}_\alpha^{-2}+\sigma^{-2}}(\frac{y}{\sigma^2}+\frac{\bar{\alpha}}{\bar{\sigma}_\alpha^2})]^2\}\exp\{-\frac{(y-\bar{\alpha})^2}{2(\sigma^2+\bar{\sigma}_\alpha^2)}\}$. Integrating α out we have $f(y) = \int f(y,\alpha)d\alpha = \frac{1}{\sqrt{2\pi(\bar{\sigma}_\alpha^{-2}+\sigma^{-2})}}\frac{1}{\sigma\bar{\sigma}_\alpha}\exp\{-\frac{(\bar{\alpha}-y)^2}{2(\sigma^2+\bar{\sigma}_\alpha^2)}\}$ and $g(\alpha|y) = \frac{f(y,\alpha)}{f(y)} = \sqrt{\frac{\bar{\sigma}_\alpha^{-2}+\sigma^{-2}}{2\pi}}exp\{-0.5(\bar{\sigma}_\alpha^{-2}+\sigma^{-2})[\alpha - \frac{1}{\bar{\sigma}_\alpha^{-2}+\sigma^{-2}}(\frac{\bar{\alpha}}{\bar{\sigma}_\alpha^2}+\frac{y}{\sigma^2})]^2\}$. Hence, $(\alpha|y) \sim \mathsf{N}(\tilde{\alpha}, (\bar{\sigma}_\alpha^{-2}+\sigma^{-2})^{-1})$, where $\tilde{\alpha}(y) = \frac{1}{\bar{\sigma}_\alpha^{-2}+\sigma^{-2}}(\frac{\bar{\alpha}}{\bar{\sigma}_\alpha^2}+\frac{y}{\sigma^2}) = y - \frac{\sigma^2}{\sigma^2+\bar{\sigma}_\alpha^2}(y-\bar{\alpha})$.

In example 9.9 the posterior mean $\tilde{\alpha}$ is a weighted average of the prior mean $\bar{\alpha}$ and of the observed y with weights given by $\frac{\sigma^2}{\bar{\sigma}_\alpha^2+\sigma^2}$ and $\frac{\bar{\sigma}_\alpha^2}{\bar{\sigma}_\alpha^2+\sigma^2}$. Hence, if $\bar{\sigma}_\alpha^2 \to 0$, sample information has no influence on the posterior while if $\bar{\sigma}_\alpha^2 \to \infty$, the posterior of α only reflects sample information (see figure 9.1).

Exercise 9.6 *Let $y_t \sim$ iid $\mathsf{N}(0,\sigma_y^2)$, where σ_y^2 unknown. A conjugate prior for σ_y^2 is obtained from the inverse-Gamma family $g(\sigma_y^2) \propto (\sigma_y^2)^{-a_1-1}exp\{-\frac{a_2}{\sigma_y^2}\}$ where a_1, a_2 are parameters. When $a_1 = 0.5\bar{\nu}$ and $a_2 = 0.5\bar{s}^2$, $\bar{s}^2\sigma_y^{-2} \sim \chi^2(\bar{\nu})$ where $\bar{\nu}$ are degrees of freedom and $\bar{s}^2$ a scale parameter. Assume T observations are available. Show that $g(\sigma_y^{-2}|y)$ is χ^2 with $(\bar{\nu}+T)$ degrees of freedom and scale equal to $(\bar{\nu}\bar{s}^2 + \sum_{t=1}^T y_t^2)$.*

Exercise 9.7 *Continuing with exercise 9.2, consider n independent draws of a two-point event; let α be the probability that a high event occurs and n_1 the number of high event episodes observed. Suppose that $g(\alpha)$ is of Beta(a_1,a_2) form, i.e. $g(\alpha) = \frac{\Gamma(a_1+a_2)}{\Gamma(a_1)\Gamma(a_2)}\alpha^{a_1-1}(1-\alpha)^{a_2-1}$ where $a_1,\ a_2 > 0$ and $\Gamma(.)$ is the Gamma-function. Show that the posterior distribution for α is Beta(a_1+n_1, a_2+n-n_1). Suppose $a_1 = a_2 = 2,\ n = 20,\ n_1 = 9$. Using the fact that if $\alpha \sim$Beta(a_1,a_2) then $\frac{a_2\alpha}{a_1(1-\alpha)}$ has an F-distribution with $(2a_1, 2a_2)$ degrees of freedom, provide an estimate of the posterior mean and of the posterior standard error of the odds ratio $\frac{\alpha}{1-\alpha}$. Show that the results of exercise 9.2 obtain if a_1 and a_2 approach zero.*

Next, we describe how conjugate priors can be employed in regression models. We do this in details as many problems can be cast into a (restricted) linear regression format.

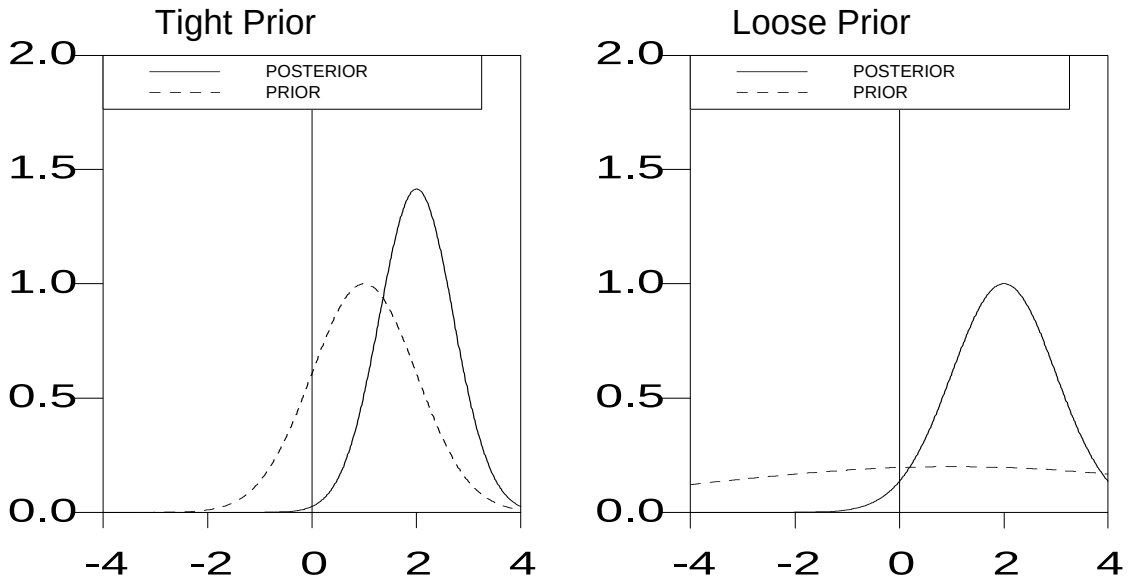


Figure 9.1: Prior and posterior densities.

Example 9.10 *Let $y_t = x_t\alpha + e_t$, where $e_t|x_t \sim \mathsf{N}(0, \sigma_e^2)$; assume $rank(x) = k$, and let $x = (x_1, \ldots x_T)'$, $y = (y_1, \ldots, y_T)$. The likelihood of y_t is: $f(y|x,\ \alpha,\ \sigma_e^2) = (2\pi)^{-0.5T}\sigma_e^{-T}exp(-0.5\sigma_e^{-2}(y - x\alpha)'(y - x\alpha))$. Assume $g(\alpha, \sigma_e^2) = g(\alpha)g(\sigma_e^2)$, let $g(\alpha) \sim \mathsf{N}(\bar{\alpha}, \bar{\Sigma}_\alpha)$ and $\bar{s}^2\sigma_e^{-2} \sim \chi^2(\bar{\nu})$. The posterior kernel is*

$$\begin{aligned}
\breve{g}(\alpha, \sigma_e^2|y, x) &= (2\pi)^{-0.5(T+k)}[2^{0.5\bar{\nu}}\Gamma(0.5\bar{\nu})]^{-1} && (9.11)\\
&\times |\bar{\Sigma}_\alpha|^{-0.5}(\bar{s}^2)^{0.5\nu}\sigma_e^{-0.5(T+\bar{\nu}+2)}exp(-0.5\bar{\nu}\bar{s}^2\sigma_e^{-2}) && (9.12)\\
&\times exp[-0.5(\sigma_e^{-2}(y - x\alpha)'(y - x\alpha) + (\alpha - \bar{\alpha})\bar{\Sigma}_\alpha^{-1}(\alpha - \bar{\alpha}))] && (9.13)
\end{aligned}$$

The exponent in (9.13) can be written as $(\alpha - \tilde{\alpha})'\tilde{\Sigma}_\alpha^{-1}(\alpha - \tilde{\alpha}) + \mathcal{Q}$ where

$$\begin{aligned}
\tilde{\Sigma}_\alpha &= (\bar{\Sigma}_\alpha^{-1} + \sigma_e^{-2}x'x)^{-1} && (9.14)\\
\tilde{\alpha} &= \tilde{\Sigma}_\alpha(\bar{\Sigma}_\alpha^{-1}\bar{\alpha} + \sigma_e^{-2}x'y) = \tilde{\Sigma}_\alpha(\bar{\Sigma}_\alpha^{-1}\bar{\alpha} + \sigma_e^{-2}x'x\alpha_{ols}) && (9.15)\\
\mathcal{Q} &= \sigma_e^{-2}y'y + \bar{\alpha}'\bar{\Sigma}_\alpha^{-1}\bar{\alpha} - \tilde{\alpha}'\tilde{\Sigma}_\alpha^{-1}\tilde{\alpha} && (9.16)
\end{aligned}$$

and $\alpha_{ols} = (x'x)^{-1}(x'y)$. Conditioning on σ_e^2 we have: $g(\alpha|\sigma_e^{-2}, y, x) \propto \exp\{-0.5(\alpha - \tilde{\alpha})\tilde{\Sigma}_\alpha^{-1}(\alpha - \tilde{\alpha})\}$ so that $(\alpha|\sigma_e^{-2}, y, x) \sim \mathsf{N}(\tilde{\alpha}, \tilde{\Sigma}_\alpha)$. Conditioning on α we have: $g(\sigma_e^{-2}|\alpha, y, x) \propto \sigma_e^{-(T+\bar{\nu}+2)}exp(-0.5\sigma_e^{-2}(\bar{\nu}\bar{s}^2 + (y - x\alpha)'(y - x\alpha))$. Hence $((\bar{\nu}\bar{s}^2 + (y - x\alpha)'(y - x\alpha))\sigma_e^{-2}|(\alpha, y, x) \sim \chi^2(T + \bar{\nu})$.

Note that if a Gamma density was used as a prior for σ_e^{-2}, the conditional for α would have been unchanged and the conditional posterior for σ_e^{-2} would be of Gamma type. It is also important to stress that in example 9.10 we calculate *conditional* posteriors. The marginal posterior of α (integrating σ_e^{-2} out) is proportional to $(\bar{s}^2 + (\alpha - \tilde{\alpha})'\tilde{\Sigma}_\alpha^{-1}(\alpha - \tilde{\alpha}) + \mathcal{Q})^{-0.5(T+k+\bar{\nu})}$. One can recognize (see e.g. the Appendix) that this is the kernel of a t-distribution with parameters $(\tilde{\alpha}, \frac{(\bar{s}^2+\mathcal{Q})\tilde{\Sigma}_\alpha}{T+\bar{\nu}}, (T+\bar{\nu}))$.

We consider two useful variants of the linear regression model in the next exercises.

Exercise 9.8 *Consider the model of example 9.10 where σ_e^2 is held fixed and $\alpha \sim \mathsf{N}(\bar{\alpha}, \bar{\Sigma}_\alpha)$. Show the form of $g(\alpha|\sigma_e^2, y, x)$ in this case (Hint: It is still normal).*

Exercise 9.9 *Suppose that the joint prior for α and σ_e^2 is non-informative i.e. $g(\alpha, \sigma_e^2) \propto \sigma_e^{-2}$. Show the form of the conditional posterior distribution for α. Is it still true that the $g(\alpha|y, x)$ is t-distributed? What are the parameters of this distribution? Show that $g(\sigma_e^2|y, x)$ is proportional to $\sigma_e^{-T-2} exp(\frac{(T-2)s^2}{2\sigma_e^2})$. Conclude that the marginal posterior for σ_e^2 is of Gamma type. Find the parameters of this distribution.*

Exercise 9.9 shows that the posterior can be proper even when the prior is not. This outcome occurs, in general, when the information content of the likelihood dominates the one of the prior.

9.2 Decision Theory

Bayesian decision theory is voluminous and too vast to be discussed here. Since some inferential decisions are based on such a theory and since Bayesian decision theory differs from classical one, we sketch the basic ideas needed to understand what will come next.

Suppose a policymaker has data on $y = (y_1, \ldots, y_T)$ and she needs to either (i) forecast y, potentially conditioning on policy intervention (say, interest rate decision); or (ii) choose an interest rate policy which maximizes consumer welfare. With each decision $d(y)$, there is an associated loss function $\mathsf{L}(\alpha, d)$ where α describes how the economy reacts to $d(y)$.

How would a frequentist approach the problem of selecting d? She would take α as given and treat d as random. Then the risk of an action is $\mathsf{R} = E\mathsf{L}(\hat{\alpha}, d)$, where the expectation is taken over decisions and d is selected to minimize risk, given some $\alpha = \hat{\alpha}$ obtained from the data. Hence, the frequentist risk is calculated conditional on $\hat{\alpha}$, averaging overall possible histories y that could have been observed as a function of d.

For a Bayesian, the history y is given and α is a random variable. Since α is random, one could choose a decision d to minimize the risk where expectations are taken with respect to α or, if a robust approach is preferred, minimize the loss obtained with the worst possible outcome i.e. $\mathsf{R} = \inf_{d \in D} \sup_{\alpha \in A} E_\alpha \mathsf{L}(\alpha, d|y^T)$. Such a decision problem is sensible when the ranking of decisions is not uniform across α.

Example 9.11 *Consider the following three equations model :*
• *Phillips curve:* $GDP_t = \alpha\pi_t + e_{s,t} \quad e_{s,t} \sim (0, 1)$

• *Demand:* $\pi_t = \Delta m_t + e_{d,t} \quad e_{d,t} \sim (0,1)$
• *Policy:* $\Delta m_t = de_{s,t}$
where variables are in deviation from steady states and suppose that the welfare function is $\mathsf{W}_t = \pi_t^2 + GDP_t^2 = (d^2 + (1+ad)^2)e_{s,t}^2 + (1+\alpha^2)e_{d,t}^2$ *so that expected welfare is* $\int \mathsf{W}_t f(e_{d,t}, e_{s,t}) de_{d,t} de_{s,t} = 1 + d^2 + (1+\alpha d)^2 + \alpha^2$. *How would an optimizing policymaker choose* d*?*

A frequentist would estimate α *from the data and minimize* $E\mathsf{L}(\hat{\alpha}, d) = 1 + \hat{\alpha}^2 + (1+\hat{\alpha})E(d(y))^2 + E(d(y))^2$*, given* $\hat{\alpha}$*. This amounts to averaging the outcomes over all possible past trajectories that could have been generated by* d*. It is immediate to see that the solution to the problem is* $d_{ML}(y) = -\frac{1}{1+\hat{\alpha}_{ML}^2}$*. A Bayesian, instead, would treat* α *as a random variable and minimize* $E\mathsf{L}(\alpha, d) = 1 + d^2 + (1 + dE(\alpha|y))^2 + E(\alpha|y)^2$*, that is, she would average the outcomes over* α*, given the observed* y*. The solution to the problem is* $d_{Bayes}(Y^T) = -\frac{1}{1+E(\alpha^2|y)}$.

One advantage of a Bayesian approach to decision making is that the maximand automatically takes into account parameter (model) uncertainty. In fact, in example 9.11 the expectation is taken with respect to the posterior $g(\alpha|y) \propto \mathcal{L}(\alpha|y)g(\alpha)$.

In general, decisions in a Bayesian framework are based on the so-called likelihood principle. This principle states that all information about the unknown α is contained in the likelihood, given the data. Hence, two likelihoods for α (from the same or different experiments) contain the same information about α if and only if they are proportional to each other. Note that the likelihood principle underlies the selection of ML-II priors.

9.3 Inference

Bayesian inference is easy since $g(\alpha|y)$ contains all the information one may need. We characterize the scope of Bayesian inference as computing $E(h(\alpha)) = \int h(\alpha) dg(\alpha|y)$ where h is a continuous function of α. Examples that fit in this characterization are numerous. For instance, $h(\alpha)$ could represent posterior moments or posterior quantiles of α; the difference in the loss function corresponding to two actions e.g. $h(\alpha) = \mathsf{L}(\alpha, d_1) - \mathsf{L}(\alpha, d_2)$; or restrictions on α, i.e. $h(\alpha) = \mathcal{I}_{A_1}(\alpha)$ where A_1 is a set and $\mathcal{I}$ an indicator function. Alternatively, $h(\alpha)$ could represent future values of the endogenous variables, i.e. $h(\alpha) = h(y^\tau)$, where $y^\tau = (y_{t+1}, \ldots, y_{t+\tau})$ and h captures turning points, prediction intervals, etc.. Finally, it could represent posterior impulse responses, variance decompositions or other statistics which are deterministic functions of α.

Sometimes, and primarily for comparison with non-Bayesian methods, one reports a point estimate of $h(\alpha)$ and an associated measure of uncertainty. These measures are justified from a Bayesian point of view either as crude approximations to the peak and the curvature of the posterior, or as a summary of posterior information as $T \to \infty$.

Let $\mathsf{L}(\hat{\alpha}^0, \alpha^0)$ be a loss function $A \times A \to \mathsf{R}$, where $\alpha^0 = h(\alpha)$ or $\alpha^0 = [y_{t+1}, \ldots y_{t+\tau}]$,

etc. and let $\hat{\alpha}^0$ be an estimator of α^0. A Bayes point estimate $\tilde{\alpha}^0$ is obtained as

$$\tilde{\alpha}^0 = argmin_{\hat{\alpha}^0} E(\mathsf{L}(\hat{\alpha}^0, \alpha^0)|y) = argmin_{\hat{\alpha}^0} \int \mathsf{L}(\hat{\alpha}^0, \alpha^0) g(\alpha^0|y) d\alpha^0 \tag{9.17}$$

There are several loss functions one could use in (9.17). Here is a brief list of candidates:

1. Quadratic loss: $\mathsf{L}(\hat{\alpha}^0, \alpha^0) = (\hat{\alpha}^0 - \alpha^0)' W (\hat{\alpha}^0 - \alpha^0)$, where W is a positive definite weighting matrix. Then $\tilde{\alpha}^0 = E(\alpha^0|y) = \int \alpha^0 dg(\alpha^0|y)$.

2. Quantile loss: $\mathsf{L}(\hat{\alpha}^0, \alpha^0) = \mathsf{L}_1(\hat{\alpha}^0 - \alpha^0)\mathcal{I}_{[-\infty, \hat{\alpha}^0]}(\alpha^0) + \mathsf{L}_2(\alpha^0 - \hat{\alpha}^0)\mathcal{I}_{[\hat{\alpha}^0, \infty]}(\alpha^0)$, $\mathsf{L}_1, \mathsf{L}_2 > 0$. Then $\tilde{\alpha}^0 = P(\alpha^0 \leq \hat{\alpha}^0|y) = \frac{\mathsf{L}_2}{\mathsf{L}_1 + \mathsf{L}_2}$. When $\mathsf{L}_1 = \mathsf{L}_2$, $\tilde{\alpha}^0$ is the median.

3. 0/1 loss: $\mathsf{L}(\hat{\alpha}^0, \alpha^0, \epsilon) = 1 - \mathcal{I}_{\epsilon(\hat{\alpha}^0)}(\alpha^0)$, where $\epsilon(\hat{\alpha}^0)$ is an open ϵ-neighborhood of $\hat{\alpha}^0$. Since $\lim_{\epsilon \to 0} argmin\ \mathsf{L}(\hat{\alpha}^0, \alpha^0) = argmin\ g(\alpha^0|y)$, then $\tilde{\alpha}^0 = argmin\ g(\alpha^0 \in \epsilon(\hat{\alpha}^0)|y)$.

For proofs of the above statements see, e.g. Berger (1985, p. 161-162). Clearly, if the posterior is normal, the choice of loss function is irrelevant (posterior mean, median and mode coincide). Note that, if the loss is quadratic and $W = I$:

$$\begin{aligned} E[(\hat{\alpha}^0 - \alpha^0)(\hat{\alpha}^0 - \alpha^0)'|y) &= (\hat{\alpha}^0 - E(\alpha^0|y))(\hat{\alpha}^0 - E(\alpha^0|y))' \\ &+ E(E(\alpha^0|y) - \alpha^0)(E(\alpha^0|y) - \alpha^0)') \\ &= Bias + variance = MSE \end{aligned} \tag{9.18}$$

Hence $\tilde{\alpha}^0 = E(\alpha^0|y)$ minimizes the Mean Square Error (MSE) of α.

It is useful to digress for a moment and compare Classical and Bayesian point estimation procedures. In classical analysis an estimator is obtained conditional on a "true" parameter value, i.e. $\tilde{\alpha}^0 = argmin_{\hat{\alpha}^0} E(\mathsf{L}(\hat{\alpha}^0, \alpha^0)|\alpha^0) = argmin_{\hat{\alpha}^0} \int \mathsf{L}(\hat{\alpha}^0, \alpha^0)) g(y|\alpha^0) dy$. Since this expression depends on unknown α^0, the solution become a function of α^0. Suppose instead we choose an estimator to minimize:

$$\begin{aligned} \tilde{\alpha}^0 &= argmin_{\hat{\alpha}^0} E_{\alpha^0} E_y(\mathsf{L}(\hat{\alpha}^0, \alpha^0)|\alpha^0) \\ &= argmin_{\hat{\alpha}^0} \int \int \mathcal{W}(\alpha^0) \mathsf{L}(\hat{\alpha}^0, \alpha^0)) g(y|\alpha^0) dy d\alpha^0 \\ &= argmin_{\hat{\alpha}^0} \int [\int \mathsf{L}(\hat{\alpha}^0, \alpha^0)) g(\alpha^0|y) d\alpha^0] \mathcal{W}(y) dy \end{aligned} \tag{9.19}$$

where $\mathcal{W}(\alpha^0)$ is a weighting function and $g(\alpha^0|y)\mathcal{W}(y) = g(y|\alpha^0)\mathcal{W}(\alpha^0)$. The minimizer of (9.19) is the one which minimizes the expression in brackets, and this is a Bayes estimator. Hence, a specification which sets up a loss function and weights parameter values by $\mathcal{W}(\alpha^0)$ implies that the Bayes estimator is also best from a frequentist point of view.

As in classical analysis one can construct confidence intervals around point estimates.

Definition 9.2 *(Credible set) A set A^0 such that $P(\alpha^0 \in A^0 \subseteq A|y) \equiv \int_{A^0} g(\alpha^0|y) d\alpha^0 = 1 - \varrho$ is the $100(1-\varrho)\%$ credible set for α^0 with respect to $g(\alpha^0|y)$.*

A credible set measures the a-posteriori degree of beliefs that $\alpha^0 \in A^0$. A classical confidence interval $CI(y)$ satisfies $P(\alpha^0 \in CI(y)|\alpha)) \equiv \int_y [\int_{CI(y)} g(\alpha^0|y,\alpha)d\alpha^0] f(y|\alpha)dy = 1-\varrho$. Also, $CI(y)$ depends on α^0. Therefore, a confidence interval is a random variable chosen so that it covers the true parameter value with probability $1-\varrho$.

Example 9.12 *Suppose a potential manager has scored 115 points in an aptitude test, suppose that the test score* $y \sim \mathsf{N}(\alpha, 100)$*, where* α *is the "true" ability of the manager. If a-priori* $\alpha \sim \mathsf{N}(100, 225)$*, the predictive density of* y *is normal with mean 100 and variance 325. Using the logic of example 9.9, it is immediate to show that* $g(\alpha|y)$ *is normal with mean* $\frac{100*100+115*225}{100+225} = 110.39$ *and variance* $\frac{100*225}{100+225} = 69.23$*. Then a 95 percent credible set for* α *is* $[110.39 \pm (1.96)(\sqrt{69.23})] = [94.08, 126.7]$*. On the other hand, a classical 95 percent confidence interval for* α *is* $[115 \pm (1.96)10] = [95.4, 134.6]$*, which is larger than the Bayesian credible set.*

Exercise 9.10 *Suppose that the number of firms which go bankrupt every week in a region has a Pareto distribution with parameters* (a_0, a_1)*, i.e.* $f(y|a_0, a_1) = \frac{a_1}{a_0}(\frac{a_0}{y})^{a_1+1}\mathcal{I}_{(a_0,\infty)}(a)$ *for* $0 < a_0 < \infty,\ a_1 > 1$*. Suppose* a_0 *is given but that nothing is known about* a_1 *so that a non-informative prior is chosen, i.e.* $g(a_1) = a_1^{-1}\mathcal{I}_{(0,\infty)}(a_1)$*. Suppose that the last 10 weeks the number of bankruptcies observed are* $(0, 2, 5, 1, 0, 1, 3, 4, 0, 5)$*. Find a 68 percent credible set for* a_1*.*

Since credible sets may not be unique, one typically chooses the highest $100(1-\varrho)\%$ credible set, i.e. a set such that $g(\alpha^0|y) \geq \kappa(\varrho),\ \forall \alpha^0 \in A^0$, where $\kappa(\varrho)$ is the largest constant such that $P(\alpha^0 \in A^0|y) = 1-\varrho$.

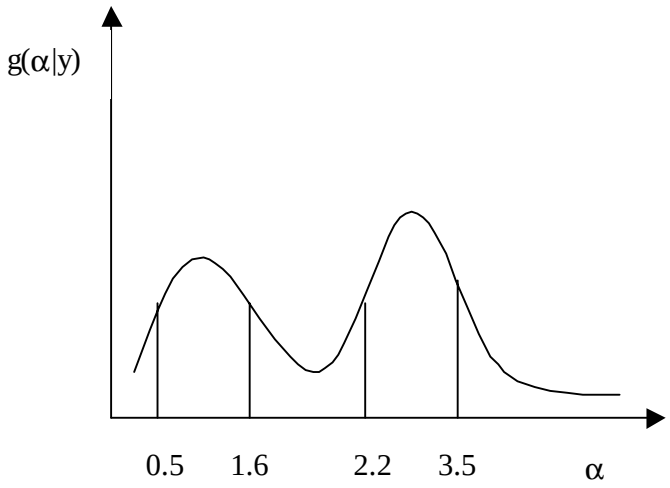


Figure 9.2: Highest credible set

Example 9.13 *Problems in computing credible sets may occur when the posterior has multiple modes. In the that case a credible set may be disjoint. For example, in figure 9.2, such a set includes the area between 0.5 and 1.6 and between 2.2 and 3.5.*

9.3.1 Inference with Multiple Models

In many situations one is faced with the dilemma of choosing a model for the analysis among a variety of alternatives. In a classical framework, one uses tests to decide which specification (e.g. the length of a VAR model) to employ. In a Bayesian framework it is optimal not to discard any model but instead appropriately weight their outcomes using their posterior probability.

Let $f(y|\alpha_j, \mathcal{M}_j)$ be the likelihood for model j; let $g(\alpha_j|\mathcal{M}_j)$ be the prior for α_j and $g(\mathcal{M}_j)$ the prior on model j, $j = 1, \ldots, J$, where $\sum_j g(\mathcal{M}_j) = 1$. Suppose we are interested in the conditional mean of $h(\alpha)$. Then

$$E[h(\alpha)] = \sum_j E[h(\alpha)|y, \mathcal{M}_j] g(\mathcal{M}_j|y) \tag{9.20}$$

The first element of (9.20) was previously calculated. Conditional on $\mathcal{M}_j$, for $\alpha_j \in A_j$:

$$E[h(\alpha)|y, \mathcal{M}_j] = \frac{\int_{A_j} h(\alpha_j) f(y|\alpha_j, \mathcal{M}_j) g(\alpha_j|\mathcal{M}_j) d\alpha_j}{\int_{A_j} f(y|\alpha_j, \mathcal{M}_j) g(\alpha_j|\mathcal{M}_j) d\alpha_j} \tag{9.21}$$

The posterior for model j is

$$\begin{aligned} g(\mathcal{M}_j|y) &= \frac{f(y|\mathcal{M}_j) g(\mathcal{M}_j)}{f(y)} = \frac{g(\mathcal{M}_j) \int f(y|\alpha_j, \mathcal{M}_j) g(\alpha_j|\mathcal{M}_j) d\alpha_j}{f(y)} \\ &\propto g(\mathcal{M}_j) \int f(y|\alpha_j, \mathcal{M}_j) g(\alpha_j|\mathcal{M}_j) d\alpha_j = g(\mathcal{M}_j) f(y|\mathcal{M}_j) \end{aligned} \tag{9.22}$$

Hence $g(\mathcal{M}_j|y)$ is the product of the prior probability of model j and its predictive density.

Therefore, to calculate $E[h(\alpha)]$ we need three steps: (i) compute the posterior expectation of $h(\alpha)$ for each model $\mathcal{M}_j$ using (9.21), (ii) obtain the predictive density and combine it with the prior $g(\mathcal{M}_j)$ as suggested in (9.22), (iii) average $E[h(\alpha|y, \mathcal{M}_j)]$ across models as suggested in (9.20).

When is it appropriate to choose only one of the J available models? It is easy to check that such a choice is appropriate when $g(\mathcal{M}_j|y)$ is independent of j and $E(h(\alpha|y, \mathcal{M}_j))$ is roughly constant across j or when $g(\mathcal{M}_j|y)$ is close to one for some j.

Example 9.14 *Suppose two commercial forecasters are producing forecasts of GDP growth one quarter ahead. Suppose that* $(y_{t+1}|\mathcal{M}_1) = 2.5$ *and that* $(y_{t+1}|\mathcal{M}_2) = 1.5$ *and that both commercial forecasters have been equally successful in the past so that* $g(\mathcal{M}_1) = g(\mathcal{M}_2) = 0.5$. *Suppose* $f(y|\mathcal{M}_1) = 0.8$ *and* $f(y|\mathcal{M}_2) = 1.2$. *Optimal model combination implies that the Bayes forecast of GDP growth is 2.5*(0.5*0.8)+1.5*(0.5*1.2)=1.0+0.9=1.9.*

9.3.2 Normal Approximations

In classical analysis one uses asymptotic approximations to derive the properties of estimators and to test hypotheses. One could take similar approximations also in a Bayesian framework. For example, when $g(\alpha|y)$ is unimodal and roughly symmetric, and the mode α^* is in the interior of A, a normal distribution centered at α^* could be used. That is:

$$\log g(\alpha|y) \approx \log g(\alpha^*|y) + 0.5(\alpha-\alpha^*)'[\frac{\partial^2 \log g(\alpha|y)}{\partial\alpha\partial\alpha'}|_{\alpha=\alpha^*}](\alpha-\alpha^*) \tag{9.23}$$

Since $\log g(\alpha^*|y)$ is a constant from the point of view of α, letting $\Sigma(\alpha^*) = -[\frac{\partial^2 \log g(\alpha|y)}{\partial\alpha^2}|_{\alpha=\alpha^*}]$, we have that $g(\alpha|y) \approx \mathsf{N}(\alpha^*, \Sigma(\alpha^*)^{-1})$ and an approximate $100(1-\varrho)\%$ highest credible set is $\alpha^* \pm \mathsf{SN}\ (\varrho/2)(\Sigma(\alpha^*)^{-0.5})$, where $\mathsf{SN}(\varrho/2)$ is the standard normal evaluated at $(\varrho/2)$.

This approximation is valid under regularity conditions when $T \to \infty$ (see below) or when the posterior kernel is roughly normal. It is highly inappropriate when:

- The likelihood is flat in some dimension ($-[\frac{\partial^2 \log g(\alpha|y)}{\partial\alpha^2}|_{\alpha=\alpha^*}]$ is poorly estimated).
- The likelihood function has multiple peaks (a single peak approximation is incorrect).
- The likelihood function is unbounded (no posterior mode exists).
- The mode is on the boundary of A (there is a natural truncation).
- $g(\alpha|y) = 0$ in the neighborhood of α^*.

Example 9.15 *Let y_t be iid $\mathsf{N}(\alpha, \sigma^2)$ and assume a non-informative prior for $(\alpha, \log\sigma)$. The joint posterior density is $g(\alpha, \log\sigma|y) \propto -T\log\sigma - \frac{0.5}{\sigma^2}[(T-1)s^2 - T(\bar{y}-\alpha)^2]$ where $\bar{y}$ is the sample mean and s^2 is the sample variance of y. Then $\frac{\partial g(\alpha,\log\sigma|y)}{\partial\alpha} = \frac{T(\bar{y}-\alpha)}{\sigma^2}$; $\frac{\partial g(\alpha,\log\sigma|y)}{\partial\log\sigma} = -T + \frac{(T-1)s^2+T(\bar{y}-\alpha)^2}{\sigma^2}$, so that the mode is $(\alpha^* = \bar{y},\ \log\sigma^* = 0.5\log\frac{(T-1)s^2}{T})$. The matrix of second derivatives with respect to $(\alpha, \log\sigma)$ evaluated at the mode is diagonal with elements equal to $-\frac{T}{\sigma^2}$ and $-2T$. Hence $g(\alpha, \log\sigma|y) \approx \mathsf{N}\left(\begin{bmatrix}\bar{y}\\ 0.5\log\frac{(T-1)s^2}{T}\end{bmatrix}, \begin{bmatrix}-\frac{T}{\sigma^2} & 0\\ 0 & -2T\end{bmatrix}^{-1}\right)$*

Exercise 9.11 *Suppose you want to study the effects of fiscal policy in a stagflation. You have data (x_i, n_i, y_i) across countries (or across experiments using DSGE models), where x_i represents the magnitude of the fiscal impulse in country i given in the $n_i - th$ experiment and y_i is the proportion of cases the economy has recovered (n_i is the number of instances in which a fiscal policy shock of a particular size has occurred). Suppose $f(y_i|\alpha_1, \alpha_2, n_i, x_i) \propto (exp\{\alpha_1 + \alpha_2 x_i\})^{y_i}(1 - exp\{\alpha_1 + \alpha_2 x_i\})^{n_i - y_i}$. Suppose the prior for (α_1, α_2) is non-informative i.e. $g(\alpha_1, \alpha_2) \propto 1$. Compute a normal approximation to the posterior of (α_1, α_2) and obtain approximate 68% and 95% contours.*

Because of the focus of this book, we only briefly describe what happens to Bayes estimators when $T \to \infty$. In classical analysis one would provide conditions for consistency and asymptotic normality of the estimators. Here there is no true parameter value toward which the estimator would converge asymptotically and which could be used as a pivot for constructing the asymptotic (normal) distribution. To have something which resembles consistency and asymptotic normality we need the following concept of information.

Definition 9.3 *Let the density of the model be $f(y|\alpha)$ and let the true density be $f^+(y)$. Let $t = 1, \ldots T$. The Kullback-Leibler (KL) information is defined at any value α by*

$$KL(\alpha) = E[\log \frac{f^+(y_t)}{f(y_t|\alpha)}] = \int \log \frac{f^+(y_t)}{f(y_t|\alpha)} f^+(y_t) dy_t \tag{9.24}$$

In words, the $KL(\alpha)$ measures the discrepancy between the model distribution and the true distribution of the data. Using KL information one could define consistency as follows

Result 9.1 *(Consistency) Suppose data is modelled with a parametric distribution function $f(y|\alpha)$ and a prior $g(\alpha)$ is assumed. Suppose that true data density belongs to $f(y|\alpha)$, i.e. $f^+ = f(y|\alpha_0)$ for some α_0. Then, as $T \to \infty$, $\alpha^* \xrightarrow{P} \alpha_0$.*

Example 9.16 *Suppose $y_t = x_t\alpha + e_t$, $e_t \sim \mathsf{N}(0,1)$. Assume $\frac{1}{T}\sum_t x_t'x_t \xrightarrow{P} \Sigma_{xx}$, $\frac{1}{T}\sum_t x_t'e_t \xrightarrow{P} 0$ and that $\alpha \sim \mathsf{N}(0, \sigma_\alpha^2 I)$. The posterior mean of α is $\tilde{\alpha} = (x'x + \sigma_a^{-2}I)^{-1}x'y$ where $x = (x_1, \ldots x_t)'$ and $y = (y_1, \ldots y_t)'$. As $T \to \infty$ $\tilde{\alpha} \to \alpha_{OLS}$. Moreover, $\tilde{\alpha} = \alpha + (x'x + \sigma_a^{-2}I)^{-1}x'e = \alpha + (\frac{1}{T}\sum_{t=1}^T x_t'x_t + \frac{1}{T\sigma_a^2}I)^{-1}\frac{1}{T}\sum_{t=1}^T x_t'e_t \to \alpha$ as $T \to \infty$ so that $\tilde{\alpha}$ is consistent.*

When the true data density is not included in the parametric family there is no longer a true α_0. If α_0 is the minimizer of the KL information, then consistency can still be proved (see e.g. Bauwens, Lubrano and Richard (1999)).

Result 9.2 *(Asymptotic Normality) Suppose α_0 is not on the boundary of the parameter space. If $\alpha^* \xrightarrow{P} \alpha_0$ as $T \to \infty$, $g(\alpha|y) \to \mathsf{N}(\alpha_0, (T\Sigma(\alpha_0))^{-1})$ where $\Sigma(\alpha_0) = -E[\frac{\partial^2 \log g(\alpha|y)}{\partial\alpha\partial\alpha'}|_{\alpha=\alpha_0}]$. Here $\Sigma(\alpha_0)$ can be estimated using $\Sigma(\alpha^*)$ and α_0 either satisfies $f^+ = f(y|\alpha_0)$ or is the minimizer of (9.24).*

9.3.3 Testing hypotheses/relative fit of different models

Testing hypotheses or evaluating models in a Bayesian framework means calculating their relative posterior support. One simple way of evaluating alternatives is to use the posterior odds (P0) ratio

$$PO = \frac{g(\mathcal{M}_j|y)}{g(\mathcal{M}_{j'}|y)} = \frac{g(\mathcal{M}_j)}{g(\mathcal{M}_{j'})} \times \frac{f(y|\mathcal{M}_j)}{f(y|\mathcal{M}_{j'})} \tag{9.25}$$

In (9.25) the first term is the prior odds, the second is the Bayes factor.

Example 9.17 *Suppose you are betting on the stability of a fixed exchange rate regime. Suppose that under the null hypothesis (say, normal conditions) there is a 50-50 chance that the regime will be maintained. Under the alternative hypothesis (say, increasing oil prices) the probability that the fixed exchange rate regime will be maintained is 0.25. Suppose that a-priori both hypotheses are equally probable and that the fixed exchange rate regime has been maintained in 90 of the 100 months for which you have data. Then*

$$PO = \frac{0.5}{0.5} \times \frac{(0.5)^{0.1}(0.5)^{0.9}}{(0.75)^{0.1}(0.25)^{0.9}} = \frac{0.5}{0.2790} = 1.79 \tag{9.26}$$

Hence, 90 months of fixed exchange rates have changed the odds of the null from 1 to 1.79.

Exercise 9.12 *Continuing with example 9.12 suppose you are interested in classifying managers as been below or above average. Let $\mathcal{M}_0 : \alpha \leq 100$; $\mathcal{M}_1 : \alpha > 100$. Find the posterior odds ratio for $\mathcal{M}_0$ vs. $\mathcal{M}_1$.*

As it is clear from example 9.17, the Bayes factor is the ratio of the predictive densities of the two models, i.e $f(y|\mathcal{M}_j) = \int_R f(y|\mathcal{M}_j, \alpha_j)g(\alpha_j)d\alpha_j$. Predictive densities can also be interpreted as predictive scores (Kass and Raftery (1995)). In fact, as likelihood functions, they can be decomposed into the product of densities of one-step ahead prediction errors (see later on). Hence, they inform us on the relative fit of the two models to the data.

As we have done with Bayes estimators, it is possible to derive the asymptotic properties of Bayes factors. The interested reader may consult Kass and Raftery (1995). Roughly speaking, Bayes factors provide a consistent model choice selection criteria when (a) the posterior distribution asymptotically concentrates around the pseudo ML estimator, (b) the pseudo ML converges in probability to the pseudo true value, (c) Bayes factor chooses the model which is closest to the pseudo true model in a Kullback-Leibler sense.

A Bayes factor differs from a likelihood ratio statistics: in fact, the relative agreement of prior and likelihood and the least square fit of the models matter for the selection.

Example 9.18 *Let $y_j = x_j\alpha_j + e_j$, $e_j \sim \mathsf{N}(0, \sigma_j^2)$, $j = 1, 2$. Suppose $g(\alpha_j) \sim \mathsf{N}(\bar{\alpha}_j, (\sigma_j^2\bar{\Sigma}_{\alpha_j})^{-1})$; $\bar{s}_j^2\sigma_j^{-2} \sim \chi^2(\bar{\nu}_j)$. If $\bar{\nu}_1 = \bar{\nu}_2 = \bar{\nu}$ and $\bar{s}_1^2 = \bar{s}_2^2$, the Bayes factor of model 1 relative to model 2 is $(\frac{|\bar{\Sigma}_{\alpha_1}||\tilde{\Sigma}_{\alpha_2}|}{|\bar{\Sigma}_{\alpha_2}||\tilde{\Sigma}_{\alpha_1}|})^{0.5}(\frac{\nu_1 s_1^2 + (\alpha_{1,ols}-\tilde{\alpha}_1)'X_1X_1(\alpha_{1,ols}-\tilde{\alpha}_1) + (\bar{\alpha}_1-\tilde{\alpha}_1)'\bar{\Sigma}_{\alpha_1}(\bar{\alpha}_1-\tilde{\alpha}_1)}{\nu_2 s_2^2 + (\alpha_{2,ols}-\tilde{\alpha}_2)'X_2X_2(\alpha_{2ols}-\tilde{\alpha}_2) + (\bar{\alpha}_2-\tilde{\alpha}_2)'\bar{\Sigma}_{\alpha_2}(\bar{\alpha}_2-\tilde{\alpha}_2)})^{-0.5(T+v)}$ where $\alpha_{jols} = (x_j'x_j)^{-1}(x_j'y_j)$; $s_j^2 = (y_j - x_j\alpha_{j,ols})^2$. The likelihood ratio statistics is $(\frac{v_1 s_1^2}{v_2 s_2^2})^{-0.5T}$.*

Predictive densities are typically hard to compute analytically since they require multi-dimensional integration. Two approximations are available in the literature.

Laplace approximation

When the likelihood is highly peaked around the mode and close to symmetric, the posterior density can be quadratically approximated around the mode. Let $f(y|\mathcal{M}_j,\ \alpha_j)g(\alpha_j|\mathcal{M}_j) \equiv$

$exp[g^{\ddagger}(\alpha_j)]$. Then $g^{\ddagger}(\alpha_j) \approx g^{\ddagger}(\alpha_j^*) + 0.5(\alpha_j - \alpha_j^*)'\Sigma_j(\alpha_j^*)(\alpha_j - \alpha_j^*)$ where the remainder is $o(||\alpha_j - \alpha_j^*||^2)$ and $\Sigma(\alpha_j) = \frac{\partial^2 g^{\ddagger}(\alpha)}{\partial\alpha\partial\alpha'}$. Integrating with respect to α we have $f^*(y|\mathcal{M}_j) = (2\pi)^{0.5k_j}|-\Sigma(\alpha_j^*)|^{-0.5}exp[g^{\ddagger}(\alpha_j^*)]$, where k_j is the dimension of α_j, so that the approximate Bayes factor is $\frac{f^*(y|\mathcal{M}_j)}{f^*(y|\mathcal{M}_{j'})} = \frac{\exp[g^{\ddagger}(\alpha_j^*)](2\pi)^{0.5k_j}|-\Sigma(\alpha_j^*)|^{-0.5}}{\exp[g^{\ddagger}(\alpha_{j'}^*)](2\pi)^{0.5k_{j'}}|-\Sigma(\alpha_{j'}^*)|^{-0.5}}$

Exercise 9.13 *Show that* $2 * \ln PO \approx 2[\ln f(\alpha_j^*|y) - \ln f(\alpha_{j'}^*|y)] - (k_j - k_{j'})\ln T +$ $(k_j - k_{j'})\ln(2\pi) + 2\ln[g(\alpha_j|\mathcal{M}_j)] - 2\ln[g(\alpha_{j'}|\mathcal{M}_{j'})] + 2\ln[g(\mathcal{M}_j) - g(\mathcal{M}_{j'})] + \ln(|-T^{-1}\Sigma(\alpha_j^*)|) - \ln(|-T^{-1}\Sigma(\alpha_{j'}^*)|)$.

Exercise 9.14 *Show the form of* $2*\ln PO$ *when* $\mathcal{M}_j$ *and* $\mathcal{M}_{j'}$ *are nested (i.e.* $\alpha_j = (\alpha_{j'}, a)$*).*

Exercise 9.13 shows that a Laplace approximation to the PO ratio can be composed in several parts: the first term is the likelihood ratio statistic (evaluated at the mode) with $k_j - k_{j'}$ degrees of freedom. The second measures the relative dimensions of the two models. This makes the Laplace approximation consistent under both the null and the alternative. The last two terms represent a correction due to the estimated curvature at the mode of the two models. Since for $T \to \infty$ they disappear, they represent a small sample correction to the adjusted ML ratio statistic.

Schwarz approximation

The Laplace approximation to the PO ratio requires the specification of a $g(\alpha)$. The Schwarz approximation does not. However, while the error in the Laplace approximation is $O(T^{-1})$, in the Schwarz approximation it is $O(1)$; that is, it is independent of the sample size. The Schwarz approximation is

$$SCA = \log[f(y|\mathcal{M}_j, \alpha_{j,ML})] - \log[f(y|\mathcal{M}_{j'}, \alpha_{j',ML})] - 0.5(k_j - k_{j'})\ln(T) \tag{9.27}$$

where $\alpha_{j,ML}$ is the maximum likelihood estimator of α in model j. It is easy to see that SCA uses the first three terms of the Laplace approximation to $2lnPO$ but evaluates them at α_{ML} instead of at α^*. Note also that, as $T \to \infty$, $\frac{SCA - \log PO}{\log PO} \to 0$.

Testing a point null is difficult in a Bayesian framework since a continuous prior on A implies that $g(\alpha_0) = 0$. There are two routes one can take. First, since a point null is a restriction on an interval around α_0, we could consider a prior on $(\alpha_0 \pm \epsilon)$, where ϵ is small relative to the posterior standard deviation of α. This would be the case, for example, when the likelihood is flat over $\alpha_0 \pm \epsilon$. In this situation, the PO ratio is well defined.

Alternatively, a prior mixing discrete and continuous distributions could be specified, i.e. $g(\alpha_0) = g_0$ and $g(\alpha \neq \alpha_0) = (1 - g_0)g_1(\alpha)$, where $g_1(\alpha)$ is a proper prior. Examples of this specification appear, for example, in Bayesian testing of unit roots (see e.g. Sims (1988)). There, a discrete prior is given to the unit root and a mixed discrete-continuous prior to the stationary region.

A question that often arises in practice is what to do when we need to compare several models, not just two. In that case Leamer's measure of posterior probability becomes useful. Such a measure is given by:

$$LEA(\mathcal{M}_j|y) = \frac{g(\mathcal{M}_j|y)PO_{j0}}{\sum_{j'} g(\mathcal{M}_{j'}|y)PO_{j'0}} \tag{9.28}$$

When the set of possible models is large, one should be careful in assuming equal a-priori probability on each of them since such a choice may counterintuitively assigns a large weight on models which are large in size.

Example 9.19 *(Sala, Doppelhofer, Miller) Suppose you have a large number of possible determinants of growth and you are interested in examining what is the posterior probability that a variable is important for growth, where models here are characterized by combinations of the potential explanatory variables. It is easy to verify that if there are k possible regressors, the number of possible models is 2^k. If an equal prior probability of $\frac{1}{2^k}$ is used on each model, the expected model size is $\frac{k}{2}$. This means that if k=20, the a-priori expected number of regressors is 10.*

In such a situation it is better to select the prior mean for the model size and let each regressor have prior probability equal to $1/k$ times this prior mean.

Exercise 9.15 *Consider the problem of forecasting quarterly exchange rate changes and suppose you have five possible candidate variables: a constant, the price differential, the interest rate differential, the output differential and the money differential (therefore, there are 32 possible model specifications). Using the dollar-yen exchange rate and data on prices, output, interest rates and money for the US and Japan, compute the posterior mean and the posterior standard deviation for each regressor and the posterior probability that each regressor is zero (i.e. compute one minus the sum of posterior probabilities of the models where that variable appear). What is the posterior probability that the best model for the dollar-yen exchange rate is a random walk with drift?*

9.3.4 Forecasting

Forecasting is straightforward in a Bayesian framework since, as we have seen, the problem fits well into the calculation of $E(h(\alpha))$. The predictive density for future y's in model j is:

$$f(y_{t+1},\ldots,y_{t+\tau}|y_t,\ldots,y_1,\mathcal{M}_j) = \int g(\alpha_j|y_t,\mathcal{M}_j)\prod_{i=t+1}^{t+\tau} f(y_i|y_{i-1},\alpha_j,\mathcal{M}_j)d\alpha_j \tag{9.29}$$

The first term in (9.29) is the posterior of α, conditional on model j and the second term is the recursive one-step ahead predictive density constructed from the model.

Example 9.20 *Let $y_t = x_t\alpha + e_t, \quad e_t \sim \mathsf{N}(0, \sigma_e^2)$. Suppose σ_e^2 fixed, let $g(\alpha) \sim \mathsf{N}(\bar{\alpha}, \bar{\Sigma}_\alpha)$ and let $x_t^\tau = [x_{t+1}, \dots x_{t+\tau}]$ be known. Since $g(\alpha|y) \sim \mathsf{N}(\tilde{\alpha}, \tilde{\Sigma}_\alpha)$ and since $(y_t^\tau|\alpha, y_t, x_t, x_t^\tau) \sim \mathsf{N}(x_t^\tau\alpha, \sigma_e^2 I)$, we have that $(y_t^\tau|y_t, x_t, x_t^\tau) \sim \mathsf{N}(x_t^\tau\tilde{\alpha},\ x_t^\tau\tilde{\Sigma}_\alpha x_t^\tau + \sigma_e^2 I)$.*

Exercise 9.16 *Using the same setup of example 9.20 show that, if $\bar{s}^2\sigma_e^{-2} \sim \chi^2(\bar{\nu})$, $(y_t^\tau|y_t, x_t, x_t^\tau)$ has a t-distribution. Show the parameters of this distribution.*

If one is interested in choosing the best forecasting model (the model which has the highest posterior support), and two alternatives are available, one can use the predictive odds ratio which is given by:

$$POR = \frac{g(\mathcal{M}_j)}{g(\mathcal{M}_{j'})} \frac{f(y_{t+1}, \dots, y_{t+\tau}|y_t, \dots, y_1, \mathcal{M}_j)}{f(y_{t+1}, \dots, y_{t+\tau}|y_t, \dots, y_1, \mathcal{M}_{j'})} \tag{9.30}$$

Note that each $f(y_i, |y_{i-1}, \alpha_j, \mathcal{M}_j)$ in (9.29) is a measure of the density of the one-step ahead error made in predicting y_i, given y_{i-1}. Therefore, examining model adequacy (as described by the predictive density) is the same as checking its one-step ahead out-of-sample forecasting performance.

9.4 Hierarchical and Empirical Bayes models

Hierarchical structures are useful to model situations where repeated observations on the same phenomena are available or when either the prior or the likelihood can be broken down into stages. For example, one may guess that parameter estimates obtained in different experiments may be connected (e.g. learning about rationality using experiments on different groups of individuals). In other cases, parameters may come in two layers and at one level there is some information while, at the other, little is known (e.g. there is some knowledge about the evolution of the parameters of a Phillips' curve over time but little is known about the distribution of the parameters regulating its evolution). Finally, there could be latent variables and a parametric model describing how latent variables are generated is available (e.g. in Arbitrage pricing (APT) models).

Consider first the case of a prior density with two stages: $g(\alpha, \theta) = g(\alpha|\theta)g(\theta)$ where θ is a vector of *hyperparameters*. The joint posterior is $g(\alpha, \theta|y) \propto f(y|\alpha, \theta)g(\alpha|\theta)g(\theta)$; and the marginal posteriors are $g(\alpha|y) = \int_{\mathbb{R}} g(\alpha, \theta|y)d\theta$ and $g(\theta|y) = \int_{\mathbb{R}} g(\alpha, \theta|y)d\alpha$.

The case when the likelihood has two stages can be similarly handled. Let $f(y|z, \alpha, \theta) = f(y|z, \alpha)f(z|\theta)$. If $g(\alpha, \theta)$ is the prior, $g(z, \alpha, \theta) = f(z|\theta)g(\alpha, \theta)$ is the joint prior and the joint posterior is $g(\alpha, \theta, z|y) \propto f(y|z, \alpha)g(z, \alpha, \theta)$. Then, the marginal posterior for the latent variable is $g(z|y) = \int_{\mathbb{R}} g(\alpha, \theta, z|y)d\alpha d\theta$ and the marginal posterior for α or θ can be similarly computed. Hence, a latent variable model is a hierarchical model with a two-stage hierarchy.

This result is important: missing data, signal extraction or any problem involving unobservable variables can be handled with the same latent variable setup.

Example 9.21 *(Experimental data) Suppose you have experimental data for different groups of individuals at different points in time. Suppose each experiment is characterized by the vector (α_j, y_{ij}, n_j), where α_j represents some interesting parameter (e.g. the proportion of individuals who are rational), y_{ij} is the data generated for individual i participating in experiment j and n_j is the number of individuals in experiment j. Under some conditions, it may be reasonable to assume that α_j are drawn from the same distribution or that there are groups of individuals with the same distribution. As we will see below, we can model this dependence with a hierarchical Bayes model.*

Example 9.22 *(Probit model) Suppose we have T independent observations on y_t, each being Bernoulli distributed with $P(y_t = 1) = \mathsf{N}(x_t\alpha)$, where N is the normal distribution. For example, we have collected recession dates and $P(y_t = 1)$ is the probability of a recession at t. The model can be rewritten as $z_t = x_t\alpha + e_t$, $e_t \sim \mathsf{N}(0, \sigma_e^2)$ and $y_t = \mathcal{I}_{[z_t>0]}$, where $\mathcal{I}_.$ is an indicator function. Here z_t is a latent variable and the likelihood of y_t has two stages.*

At times it is hard to distinguish the prior from the model as the next example shows.

Example 9.23 *(Panel data) Let $y_{it} = \alpha_i + e_{it}$, $e_{it} \sim \mathsf{N}(0, \sigma_e^2)$. Assume $\alpha_i \sim \mathsf{N}(\bar{\alpha}, \bar{\sigma}_\alpha^2)$; $\bar{\alpha} \sim \mathsf{N}(\bar{\alpha}_0, \bar{\sigma}_0^2)$ and let $\bar{\sigma}_\alpha^2, \bar{\sigma}_0^2$ be fixed. These assumptions imply:*

$$\alpha_i = \bar{\alpha} + v_{1i} \quad v_{1i} \sim \mathsf{N}(0, \bar{\sigma}_\alpha^2) \tag{9.31}$$

$$\bar{\alpha} = \bar{\alpha}_0 + v_2 \quad v_2 \sim \mathsf{N}(0, \sigma_0^2) \tag{9.32}$$

So $\alpha_i = \bar{\alpha}_0 + v_2 + v_{1i}$ and $y_{it} = \bar{\alpha}_0 + v_2 + v_{1i} + e_{it}$. Here α_i could be a latent variable and (9.32) the prior. Alternatively, (9.31)-(9.32) are two stages of the hierarchical prior for α_i.

A natural way to model the dependence of parameters in experimental data or in panels is the notion of exchangeability.

Definition 9.4 *Consider $j = 1, \ldots J$ experiments (observations on different individuals or units) for which $f(y_j|\alpha_j)$ is available. If only y_j is available to distinguish the α_j and no ordering or grouping can be made, α_j must be a-priori similar. Then $(\alpha_1, \ldots, \alpha_J)$ are exchangeable if $g(\alpha_1, \ldots, \alpha_J)$ is invariant to permutations of the order of the $\alpha_j's$.*

One way to represent an exchangeable prior for α is to set $g(\alpha|\theta) = \prod_j g(\alpha_j|\theta)$, i.e. α_j are independent draws from a distribution with parameter θ. Then, the marginal prior of α is a mixture of iid distributions with weights given by $g(\theta)$ i.e. $g(\alpha) = \int \prod_j g(\alpha|\theta)g(\theta)d\theta$.

In certain finance applications the prior may depend on observables (e.g. in a CAPM model where the return on a market portfolio depends on macroeconomic variables). In this case an exchangeable prior is $g(\alpha_1, \ldots, \alpha_j|x_1, \ldots x_j) = \int \prod_j g(\alpha_j|\theta, x_j)g(\theta|x_j)d\theta$.

The next example describes when the exchangeability assumption is appropriate.

Example 9.24 *Suppose you are interested in estimating inflation rates in the Euro area. Suppose you sample inflation rates in five countries and obtain 1.7, 1.0, 0.9, 3.0, 1.8. Call these* $y_1, \ldots, y_5$. *What can we say about a potentialy*$_6$*, having observed* $y_1, \ldots, y_5$*?*
i) If there is no information to distinguish one country from the others, observed inflation rates are exchangeable and, lacking information about the time series pattern of inflation rates, the prior for $y_1, \ldots y_6$ *should be non-informative. Based on the observed sample, one could guess a posterior density for* y_6 *with mean around 1.9 and range, say, [0.4, 3.5].*
ii) Suppose you have the additional information that the six states are Ireland, Spain, Germany, the Netherlands, France, and Belgium, but that their order is random (so you cannot say which country corresponds to which number). The five inflation rates can still be treated as exchangeable but the posterior moments of y_6 *should change since Ireland and Spain had higher rates than France and Germany in the past (and only one high inflation rate has been sampled).*
iii) Suppose that you know that the sixth state is Spain. Now exchangeability is inappropriate - you have information that distinguish Spain from the other countries - and one can guess that the posterior of y_6 *will be high concentrated around 3.0.*

Note that for experiments conducted at different times, with different agents and in different laboratories it may still be reasonable to use exchangeability since these differences imply different outcomes and not necessarily different a-priori distributions.

Posterior analysis with hierarchical models is simple and exploits the version of Bayes theorem with nuisance parameters described in section 9.1. For example, $g(\alpha, \theta|y)$ is proportional to $f(y|\alpha, \theta)g(\alpha|\theta)g(\theta) = f(y|\alpha)g(\alpha|\theta)g(\theta)$. Similarly, predictive distributions can be easily computed. In hierarchical models we distinguish between two types of predictive distributions for (future) y^τ: those conditional on $\tilde{\alpha}$, a posterior estimate of α, and those conditional on α^l, where α^l is a draw from $g(\alpha|\tilde{\theta}, y)$ and $\tilde{\theta}$ is a posterior estimate of θ.

To simulate samples for the unknowns from the posterior distribution one would set up the likelihood function $f(y|\alpha)$, the priors $g(\alpha|\theta),\ g(\theta)$ and proceed as follows:

Algorithm 9.2

1) *Compute the posterior kernel* $\breve{g}(\alpha, \theta|y)$.

2) *Compute (analytically)* $g(\alpha|\theta, y)$ *(for fixed* y*, this is a function of* θ*).*

3) *Compute* $g(\theta|y)$ *either as* $g(\theta|y) = \int g(\alpha, \theta|y)d\alpha$ *or as* $g(\theta|y) = \frac{g(\alpha,\theta|y)}{g(\alpha|\theta,y)}$.

4) *Draw* θ^l *from* $g(\theta|y)$ *and* α^l *from* $g(\alpha|\theta^l, y)$ *(If* α_j *is exchangeable, draw* α_j*,* $g(\alpha_j|\theta^l, y)$ *for each* j*). Draw* y^l_τ *from* $f(y_\tau|\tilde{\alpha})$ *or from* $f(y_\tau|\alpha^l)$.

5) *Repeat step 4) L times and compute* $h(\alpha^l_j) = h(y^l_\tau|\alpha^l_j)$ *at each step. If draws are iid, estimate* $E[h(\alpha|y)]$ *via* $E[h(\alpha|y)] = \lim_{L\to\infty} \frac{1}{L} \sum_l h(\alpha^l)$.

Note that step 3) is easy if $g(\alpha|\theta)$ is conditionally conjugate. $h(\alpha)$ could include, as usual, functions of economic interest (impulse responses, welfare costs, forecasts, etc.).

Example 9.25 *(Estimating the productivity of individual workers (random effect) and the average productivity in a plant (fixed effect)). Let y_{jt} be the number of pieces completed by worker j at hour t of the day. Suppose $(y_{jt}|\alpha_j) \sim \mathsf{N}(\alpha_j, \sigma^2)$, $j = 1, \ldots, J$; σ^2 fixed where α_j is the average productivity of the worker. Let $\bar{y}_j = \frac{1}{t_j} \sum_{t=1}^{t_j} y_{jt}$, $\sigma_j^2 = \frac{\sigma^2}{t_j}$. Then $\bar{y}_j|\alpha_j \sim \mathsf{N}(\alpha_j, \sigma_j^2)$. There are three possible estimators for α_j: i) the individual mean, $\bar{y}_j$, ii) the pooled mean, $y_p = \sum_j \frac{\bar{y}_j}{\sigma_j^2} / \sum_j \frac{1}{\sigma_j^2}$, iii) the weighted mean $\bar{y}_{wj} = \varrho_j \bar{y}_j + (1 - \varrho_j) y_p$, $\varrho_j \in [0, 1]$.*

Exercise 9.17 *Show how to use between and within variations in y_{jt} to choose between i) and ii).*

What kind of exchangeable prior would induce a researcher to choose as posterior estimator i), ii) or iii)? Estimator i) will be chosen if the prior for each α_j is independent and uniform on $[-\infty, +\infty]$; estimator ii) will be chosen if $\alpha_j = \alpha$, $\forall j$ and α is uniform on $[-\infty, +\infty]$; finally, estimator iii) will be selected if the prior on α_j is iid normal. Note that i) and ii) are special cases of iii): i) obtains if $var(\alpha_j) = \infty$, ii) obtains if $var(\alpha_j) = 0$.

Assume that σ^2 is known and let $g(\alpha_1, \ldots \alpha_j|\bar{\alpha}, \bar{\sigma}_\alpha^2) = \prod_j \mathsf{N}(\alpha_j|\bar{\alpha}, \bar{\sigma}_\alpha^2)$, where $\bar{\alpha}$ is the average productivity and $\bar{\sigma}_\alpha^2$ its dispersion across workers. Let $g(\bar{\alpha}, \bar{\sigma}_\alpha^2) = g(\bar{\alpha}|\bar{\sigma}_\alpha^2) g(\bar{\sigma}_\alpha^2) \propto g(\bar{\sigma}_\alpha^2)$ (i.e. no information about $\bar{\alpha}$ is available). Then, the joint posterior of $(\alpha_j, \bar{\alpha}, \bar{\sigma}_\alpha^2)$ is

$$g(\alpha, \bar{\sigma}_\alpha^2, \bar{\alpha}|y) \propto \prod_{j=1}^{J} \mathsf{N}(\bar{y}_j|\alpha_j, \sigma_j^2) \prod_j \mathsf{N}(\alpha_j|\bar{\alpha}, \bar{\sigma}_\alpha^2) g(\bar{\sigma}_\alpha^2, \bar{\alpha}) \tag{9.33}$$

Using the logic of example 9.9, the marginal for α_j is $g(\alpha_j|\bar{\alpha}, \bar{\sigma}_\alpha^2, y) \sim \mathsf{N}(\tilde{\alpha}_j, \tilde{\Sigma}_j)$ where $\tilde{\alpha}_j = \tilde{\Sigma}_j(\frac{\bar{y}_j}{\sigma_j^2} + \frac{\bar{\alpha}}{\bar{\sigma}_\alpha^2})$, $\tilde{\Sigma}_j = (\frac{1}{\sigma_j^2} + \frac{1}{\bar{\sigma}_\alpha^2})^{-1}$ while the marginal posterior for $\bar{\alpha}$ and $\bar{\sigma}_\alpha^2$ is

$$g(\bar{\alpha}, \bar{\sigma}_\alpha^2|y) = \int g(\alpha, \bar{\alpha}, \bar{\sigma}_\alpha^2|y, x) d\alpha \propto g(\bar{\sigma}_\alpha^2) f(y|\bar{\alpha}, \bar{\sigma}_\alpha^2) = g(\bar{\sigma}_\alpha^2) \prod_j \mathsf{N}(\bar{y}_j|\bar{\alpha}, \bar{\sigma}_\alpha^2 + \sigma_j^2) \tag{9.34}$$

which can be obtained substituting the prior into the model i.e. $y_{ij} = \bar{\alpha} + e_{ij}$, $e_{ij} \sim \mathsf{N}(0, \sigma_j^2 + \bar{\sigma}_\alpha^2)$ and using the sufficient statistic $\bar{y}_j$ to rewrite the likelihood of y_{ij}. Using (9.34) it is easy to see that the marginal of $\bar{\alpha}$, conditional on $\bar{\sigma}_\alpha^2$, is normal with mean $\tilde{\bar{\alpha}}$ and variance $\tilde{\Sigma}_{\bar{\alpha}}$ where $\tilde{\bar{\alpha}} = \tilde{\Sigma}_{\bar{\alpha}} \sum_j \frac{\bar{y}_j}{\bar{\sigma}_\alpha^2 + \sigma_j^2}$ and $\Sigma_{\bar{\alpha}} = (\sum_j \frac{1}{(\bar{\sigma}_\alpha^2 + \sigma_j^2)})^{-1}$. The marginal posterior for $\bar{\sigma}_\alpha^2$ is

$$\begin{aligned} g(\bar{\sigma}_\alpha^2|y) = \frac{g(\bar{\alpha}, \bar{\sigma}_\alpha^2|y)}{g(\bar{\alpha}|\bar{\sigma}_\alpha^2 y)} &\propto \frac{g(\bar{\sigma}_\alpha^2) \prod_j \mathsf{N}(\bar{y}_j|\bar{\alpha}, \sigma_j^2 + \bar{\sigma}_\alpha^2)}{\mathsf{N}(\bar{\alpha}|\tilde{\bar{\alpha}}, \tilde{\Sigma}_{\bar{\alpha}})} \\ &\propto \tilde{\Sigma}_{\bar{\alpha}}^{0.5} [\prod_j (\sigma_j^2 + \bar{\sigma}_\alpha^2)]^{-0.5} exp\{-\frac{(\bar{y}_j - \tilde{\bar{\alpha}})^2}{2(\sigma_j^2 + \bar{\sigma}_\alpha^2)}\} \end{aligned} \tag{9.35}$$

where the second line is obtained evaluating the likelihood function at $\tilde{\bar{\alpha}}$, when $g(\bar{\sigma}_\alpha^2)$ is non-informative. Then, a posterior 68% credible set for the average productivity is $\tilde{\bar{\alpha}} \pm \sqrt{\tilde{\Sigma}_{\bar{\alpha}}}$ and a posterior 68% credible set for the individual productivity is $\tilde{\alpha}_j \pm \sqrt{\tilde{\Sigma}_j}$.

Suppose now you want to predict the productivity of a new worker whose ability is similar to the one of existing workers. To construct predictions $y_{\tilde{j},t}$ $\tilde{j} \neq 1, \dots J$ *one could use:*

Algorithm 9.3

1) *Draw* $(\bar{\alpha}^l, (\bar{\sigma}^2_\alpha)^l)$ *from* $g(\bar{\alpha}, \bar{\sigma}^2_\alpha | y)$ *and* α^l_j *from* $g(\alpha_j | \bar{\alpha}^l, (\bar{\sigma}^2_\alpha)^l, y)$.

2) *Draw* $y^l_{\tilde{j},t}$ *from* $\mathsf{N}(\alpha^l_{\tilde{j}}, \sigma^2_j)$.

3) *Repeat steps 1.-2. L times and average* $y^l_{\tilde{j},t}$ *over* l.

Exercise 9.18 *In example 9.25, what would you do if the new worker is different from all the currently employed? What if she is similar only to a subset of current workers?*

Example 9.26 *Consider the problem of predicting financial crises and suppose that they occur when a vector of* z *variables passes a threshold* z^*. *Suppose* z_t *are unobservable but related to some observable* x_t *(e.g. liquidity of the banking system, trade balance or the state of government finances) and that we observe* $y_t = 1$ *if* $z_t \geq z^*$ *and zero otherwise. Then the model is* $z_t = \alpha x_t + e_t, \quad e_t | x_t \sim \mathsf{N}(0, \sigma^2_e)$; $y_t = \mathcal{I}_{[z^*,\infty)} z_t$ *where* σ^2_e *is known. Let* $y = [y_1, \dots, y_t]'$, $z = [z_1, \dots, z_t]'$, $x = [x_1, \dots, x_t]'$ *and* $f(y, z|x, \alpha) = f(z|x, \alpha) f(y|z)$. *Then*

$$\begin{aligned} f(y, z|x, \alpha) &= (2\pi)^{0.5T} \exp\{-0.5(z - x\alpha)'(z - x\alpha) \prod_{i=1}^{T} [y_t \mathcal{I}_{[z^*,\infty)} + (1 - y_t)\mathcal{I}_{(-\infty,z^*]}\} \\ f(y|x, \alpha) &= \int f(y, z|x, \alpha) dz = \prod_{i=1}^{T} [y_t \mathsf{N}(\alpha x) + (1 - y_t)(1 - \mathsf{N}(\alpha x))] \end{aligned} \tag{9.36}$$

where $\mathsf{N}(x\alpha)$ *is the normal distribution evaluated at* $x\alpha$. *Since* $g(\alpha, y|z, x) \propto f(y, z|x, \alpha) g(\alpha)$ *and the marginal posterior for* α *is normal with variance* $\tilde{\Sigma}_\alpha = (\bar{\Sigma}^{-1}_\alpha + \sigma^{-2}_e x'x)^{-1}$ *and mean* $\tilde{\alpha} = \tilde{\Sigma}_\alpha (\bar{\Sigma}^{-1}_\alpha \bar{\alpha} + \sigma^{-2}_e x'y)$, *where* $\bar{\alpha}$ *and* $\bar{\Sigma}_\alpha$ *are the prior mean and the prior variance of* α. *Furthermore, conditional on* (α, y, x), *the posterior for* z_t *is normal with mean* αx_t *and variance* σ^2_e *and* $z_t > z^*$ *if* $y_t = 1$ *and* $z_t \leq z^*$ *if* $y_t = 0$.

9.4.1 Empirical Bayes methods

Empirical Bayes (EB) methods attempt to reduce the costs of computing marginal posteriors in hierarchical models. They do so estimating features of the prior from the data.

In example 9.25, the posterior distribution of the individual effect α_j is obtained integrating $\bar{\alpha}$ and $\bar{\sigma}^2_\alpha$ out the joint posterior. Alternatively, one could estimate $\bar{\alpha}$ and $\bar{\sigma}^2_\alpha$, for example, $\hat{\bar{\alpha}} = \frac{1}{J_1} \sum_j^{J_1} \bar{y}_j$ and $\hat{\bar{\sigma}}^2_\alpha = T^{-1}(\frac{\sum_j (\bar{y}_j - \hat{\bar{\alpha}})^2}{J_1 - 1} - \frac{\sum_t \sum_j (y_{tj} - \bar{y}_j)^2}{T(J_1 - 1)})$, $J_1 << J$, and plug-in these estimates in the formulas for the moments of the posterior distribution. That is, instead of computing $g(\alpha|y) = \int g(\alpha, \bar{\alpha}, \bar{\sigma}^2_\alpha, \theta|y) d\bar{\alpha} d\bar{\sigma}^2_\alpha$, we calculate $g(\alpha|y, \hat{\bar{\alpha}}, \hat{\bar{\sigma}}^2_\alpha)$.

We have discussed data driven priors in section 9.1.2. As in that framework, the predictive density can be used to estimate features of the prior distribution.

Example 9.27 *Let the model for unit i of a panel be $y_{it} = \alpha_i y_{it-1} + e_t,\ t = -1, 0, 1, \ldots T$ where $\alpha_i \sim (\bar{\alpha}, \bar{\sigma}_\alpha^2)$, $\bar{\alpha} \sim (\bar{\alpha}_0, \sigma_0^2)$ and $e_t \sim (0, \sigma_e^2)$. If σ_0^2 and σ_a^2 are known (or estimable from the data), an estimator of $\bar{\alpha}_0$ is $\tilde{\bar{\alpha}}_0 = (y'_{-1}\Sigma^{-1}y_{-1})^{-1}(y'_{-1}\Sigma^{-1}y)$ where $\Sigma = (\sigma_e^2 + \bar{\sigma}_\alpha^2 y'_{-1}y_{-1} + \bar{\sigma}_0^2 y'_{-1}y_{-1})$, $y_{-1} = [y_{1-1}, \ldots, y_{n-1}]'$, and $y = [y_{10}, \ldots, y_{n0}]'$.*

There are advantages and disadvantages in using EB methods. On the one hand, computations are simpler; furthermore, priors are data driven which makes them more appealing to non-Bayesian audiences; finally, despite the fact that some parameters are estimated, the form of the posterior for α is unchanged. On the other hand, posterior estimates obtained with EB methods disregard the uncertainty present in $(\bar{\alpha}, \bar{\sigma}_\alpha^2)$. This problem can be fixed (see e.g. Morris (1983)). Another problem is that estimates of $\bar{\sigma}_\alpha^2$ may be negative; finally, while there is no problem in selecting some observations to estimate the prior in times series (use a training sample), it is unclear how to do this in cross sectional environments (which units should be used?). Hence validation techniques need to be employed to examine the robustness of the conclusions one reaches.

Exercise 9.19 *Why can estimates of $\bar{\sigma}_\alpha^2$ be negative? How would you deal with this problem?*

As we will see in chapter 10, BVARs with a Minnesota prior can be handled with EB methods. There $(\bar{\alpha}, \bar{\sigma}_\alpha^2)$ are fixed at conventional values or estimated in a training sample.

9.4.2 Meta analysis

Despite the mysterious name, Meta-analysis is relatively straightforward; in fact it tries to efficiently summarize findings from different studies on a particular topic. Questions which fit into such a framework are quite common in economics. Here is a brief list:

- Is the bank lending channel an important mechanism in transmitting monetary policy shocks? (The evidence is across countries, see e.g. Angeloni et. al. (2003)).
- Does trade increase in monetary unions? (The evidence is across studies with different samples or estimators, see e.g. Rose (2004)).
- Can financial variables predict inflation in the medium run? (The evidence may be across regimes (high/low inflation); time periods, countries, etc.)
- Do agents behave in a risk averse fashion when faced with fair bets? (The evidence comes from individuals of different age, social, cultural background, etc.)
- Does local fiscal policy affects local to union-wide prices? (The evidence comes from different countries, regimes, time periods, see e.g. Canova and Pappa (2003)).

The best way to understand how to use Meta-analysis is through an example.

Example 9.28 *Consider the question of whether monetary policy can shield economies from recessions. Suppose we have $j = 1, \ldots, J$ studies coming from different regimes or countries. For each j we have two sets of data: i) an action is undertaken in T_{0j} episodes and y_{0j} recessions are observed; ii) no action is undertaken in T_{1j} episodes and y_{1j} recessions are observed, $T_j = T_{0j} + T_{1j}$. Let the probabilities of a recession in the two cases be p_{0j} and p_{1j}. Consider $\alpha_j = \log \frac{p_{1j}/(1-p_{1j})}{p_{0j}/(1-p_{0j})}$, that is, the relative probability of a recession in the two scenarios, suppose we care about α_j, $\forall j$ (single study effect) and $\bar{\alpha}$ (average effect) and suppose no information other than $(T_{ij}, y_{ij}), i = 0, 1$ is available. A crude estimate of $(\alpha_j, \bar{\alpha})$ can be obtained by taking a normal approximation to the outcome of each experiment j, i.e. assume $\alpha_j \sim \mathsf{N}(\hat{\alpha}_j, \hat{\sigma}_j^2)$ where $\hat{\alpha}_j = \log(\frac{y_{1j}}{T_{1j}-y_{1j}}) - \log(\frac{y_{0j}}{T_{0j}-y_{0j}})$ and $\hat{\sigma}_j^2 = \frac{1}{y_{1j}} + \frac{1}{T_{1j}-y_{1j}} + \frac{1}{y_{0j}} + \frac{1}{T_{0j}-y_{0j}}$, in which case $\hat{\bar{\alpha}} = \frac{1}{j} \sum_j \hat{\alpha}_j$. Can we improve upon these estimates?*

Suppose that the J studies are compatible in some sense. Here there is some latitude regarding what compatible means. It could be that the outcomes are all drawn from the same distribution; that study j carries no information about study j'; or that no study has more information than others. In all these cases information is exchangeable.

Let $\hat{\alpha}_j$ be an estimate of α_j obtained from experiment j and consider a hierarchical structure where at the first stage the likelihood of $(\hat{\alpha}_j | \alpha_j,\ \sigma_j^2)$ is $\mathsf{N}(\alpha_j, \sigma_j^2), \sigma_j^2$ known; at the second stage the conditional prior for α_j is exchangeable and $(\alpha_j | \bar{\alpha}, \bar{\sigma}_\alpha^2) \sim \mathsf{N}(\bar{\alpha}, \bar{\sigma}_\alpha^2)$; and at the third stage the marginal for $(\bar{\alpha}, \bar{\sigma}_\alpha^2)$ is non-informative. This setup is identical to the one described in example 9.25. Hence the posteriors of α_j and $\bar{\alpha}$ can be obtained with the same techniques.

Exercise 9.20 *Consider four country studies measuring the length (in months) of a recession before and after the government started using Keynesian policies. The data is assumed to be of the same quality across time periods and is as follows:*

	Before			*After*		
	Min	*Mean*	*Max*	*Min*	*Mean*	*Max*
1	*25*	*38*	*62*	*18*	*24*	*38*
2	*26*	*29*	*37*	*19*	*21*	*25*
3	*22*	*25*	*34*	*24*	*25*	*32*
4	*27*	*32*	*40*	*21*	*33*	*37*

Using a hierarchical model where the length of a recession is assumed to be exponential with parameter α and a suitable prior for α (e.g. uniform on the positive side of the real line), provide an estimate of the difference in the mean across regimes in each study and on average. Construct a posterior 95% credible set for this difference. Is there any evidence that Keynesian policies had any effect on the length of recessions?

Exercise 9.21 *Suppose you are interested in evaluating the effects of EU agricultural funds (the so-called CAP funds) on regional growth. Suppose you have run a time series regression in each region obtaining agricultural funds with output growth on the left hand side and a number of variables controlling for the individual characteristics of the regions on the right hand side and found the following coefficients on the amount of structural funds received:*

	Region 1	Region 2	Region 3	Region 4	Region 5	Region 6	Region 7	Region 8
Estimate	28.39	7.94	-2.75	6.82	-0.64	0.63	18.01	12.16
Standard error	14.9	10.2	16.3	11.0	9.4	11.4	10.4	17.6

i) Argue about the advisability of continuing to provide structural funds to these regions.
ii) Compute pooled estimates. Would your conclusions in i) change?
iii) Using a hierarchical model where the estimates in the table play the role of $\bar{y}_j$ and the standard error the role of σ_j, assume σ_j are known, a uniform prior for $(\alpha, \log(\sigma_\alpha))$ and calculate Empirical Bayes estimates of the effect of structural funds in each region and on average. What is the value of σ_α you would condition on? i.e. what is the posterior estimate of the dispersion parameter of the effects of structural funds across regions?
iv) Simulate the posterior for $(\alpha_j, \sigma_\alpha, \alpha)$ using the hierarchical structure in iii). Display $E(\alpha_j|y, \sigma_\alpha)$ for a grid of values of σ_α between 0 and 10. Display $E(\alpha|y, \sigma_\alpha)$ and the (simulated) interquantile range for each α_j. Is it fair to say that CAP funds do not boost growth? [Hint: To make the point stronger, compute the posterior of $\max(\alpha_j)$].

9.5 Posterior simulators

As we will see in more details in the next two chapters, there is large number of problems for which the posterior distribution of α cannot be computed analytically. In others, only the kernel of the posterior is available but $f(y)$ is unknown. In the most favorable situation one can take a normal approximation to the posterior to conduct inference. In others, posterior simulators are needed. This section describes in details both approaches.

9.5.1 Normal posterior analysis

When the posterior distribution is of unknown form but suspected to be close to a normal (either because the sample is large or because of the assumptions made), it is possible to undertake posterior inference simulating sequences for the unknowns from an approximate normal distribution. To do so we need the following four steps.

Algorithm 9.4

1) *Find a measure of location (typically, the mode) of the posterior distribution. Several mode finding algorithms exist in the literature. Here are two:*

- *Conditional maximization algorithm. Choose α_0 and partition $\alpha = (\alpha_1, \alpha_2)$.*
  i) *Maximize $g(\alpha_1, \alpha_2 = \alpha_{20}|y)$ with respect to α_1. Let α_1^* the maximizer.*
  ii) *Maximize $g(\alpha_1^*, \alpha_2|y)$ with respect to α_2. Let α_2^* the maximizer.*
  iii) *Set $\alpha_{20} = \alpha_2^*$. Iterate on i) and ii) until convergence is achieved.*
  iv) *Start from any other $(\alpha_{10}, \alpha_{20})$ and check if maximum is global.*
- *Newton-type algorithm. Choose an α_0, let $LG = \log g(\alpha|y)$ or $LG = \log \breve{g}(\alpha|y)$*
  i) *Compute $LG' = \frac{\partial LG}{\partial \alpha}(\alpha_0)$; $LG'' = \frac{\partial^2 LG}{\partial \alpha \partial \alpha'}(\alpha_0)$; approximate LG quadratically.*

ii) *Set* $\alpha^l = \alpha^{l-1} - \varrho(LG''(\alpha^{l-1}|y))^{-1}(LG'(\alpha^{l-1}|y)), \quad \varrho \in (0,1), l = 1,2,\ldots.$

iii) *Iterate on i)-ii) until convergence is achieved.*

Whenever analytic derivatives are difficult to calculate one could use $LG' = \frac{LG(\alpha+\delta_i e_i|y)-LG(\alpha-\delta_i e_i|y)}{2\delta_i}$ *and* $LG'' = \frac{LG(\alpha+\delta_i e_i+\delta_j e_j|y)-LG(\alpha+\delta_i e_i-\delta_j e_j|y)}{4\delta_i\delta_j} + \frac{LG(\alpha-\delta_i e_i-\delta_j e_j|y)-LG(\alpha-\delta_i e_i+\delta_j e_j|y)}{4\delta_i\delta_j}$. *This algorithm is fast if* α_0 *is "good" and LG close to quadratic. It does not work well if* LG'' *is not positive definite.*

In both algorithms crude estimates, obtained discarding parts of model and/or the data, can be used as initial conditions. For example, in hierarchical models, one could fix $(\bar{\alpha}, \bar{\sigma}^2_\alpha)$*, construct* $g(\alpha|\bar{\sigma}^2_\alpha, \bar{\alpha}, y)$ *and use this as conjugate prior for computing the mode.*

The mode α^* *does not have a special role here; it is simply the point around which to map the shape of the posterior distribution. Therefore,*

2) *Find an analytic approximation to posterior density, centered at the mode.*

The most typical approximation is a normal one, i.e. $g(\alpha|y) \approx \mathsf{N}(\alpha^*, \Sigma_{\alpha^*})$ *where* $\Sigma_\alpha = [-LG''(\alpha^*)]^{-1}$*. When multiple modes are present, construct an approximation to each mode, and set* $g(\alpha|y) \propto \sum_i \varrho_i \mathsf{N}(\alpha^*_i, \Sigma_{\alpha^*_i})$*, where* $0 \leq \varrho_i \leq 1$*. If the modes are clearly separated and a normal approximation is chosen for each of them, it is typical to select* $\varrho_i = \breve{g}(\alpha^*_i|y)|\Sigma_{\alpha^*_i}|^{0.5}$*. If the sample is small and/or the normal approximation is inappropriate (e.g. if a parameter needs to be positive) one could use a t-distribution with small number of degrees of freedom, i.e.* $g(\alpha|y) \propto \sum_i \breve{g}(\alpha|y)[\nu + (\alpha - \alpha^*_i)'\Sigma^{-1}_{\alpha^*_i}(\alpha - \alpha^*_i)]^{-0.5*(k+v)}$*, where* k *is the dimension of* α*. When* $\nu = 1$ *we approximate the posterior with a Cauchy, a distribution with large overdispersion (very thick tails; the moments do not exist). In typical economic applications,* $\nu = 4$ *or* 5 *is appropriate.*

3) *Draw samples from the approximate posterior distribution. If draws are iid, the law of large numbers permits us to approximate* $E(h(\alpha))$ *or the posterior probability contours of* $h(\alpha)$ *with* $\frac{1}{L}\sum_l h(\alpha^l)$ *or the ordered values of* $h(\alpha^l)$*. Note that if a Laplace approximation to* $g(\alpha|y)$ *is used, then* $E(h(\alpha|y)) \approx h(\alpha^*)\breve{g}(\alpha^*|y)| - \frac{\partial^2 log(h(\alpha)\breve{g}(\alpha|y))}{\partial\alpha\partial\alpha'}|_{\alpha=\alpha^*}|^{0.5}$.

4) *Check the accuracy of the approximation by computing the Importance Ratio* $IR^l = \frac{\breve{g}(\alpha^l|y)}{g^A(\alpha^l|y)}$ *where* g^A *is the approximating distribution. If* IR^l *is roughly constant across* l*, the approximation is good. If it is not, other simulation methods are needed.*

Exercise 9.22 *Consider estimating a reduced form Phillips curve* $\pi_{t+1} = \alpha_\pi \pi_t + \alpha_{gap} gap_t + e_t$ *where* gap_t *is the difference between actual and potential output and* $e_t \sim (0, \sigma^2_e)$*. Assume that* $\alpha = (\alpha_\pi, \alpha_{gap}) \sim N(\bar{\alpha}, \bar{\Sigma}_\alpha)$ *and that* $g(\sigma^2_e)$ *is non-informative. Derive the marginal posterior for* α*. Using US data on CPI inflation and linearly detrended GDP as a proxy for the gap, construct a posterior normal approximation and report a 68% credible set for* α_{gap}*.*

9.5.2 Basic Posterior Simulators

When a normal or t-distribution are not necessarily suited to approximate $g(\alpha|y)$, other posterior simulators can be used. The next two work well when IR^l is approximately constant across l.

Acceptance sampling

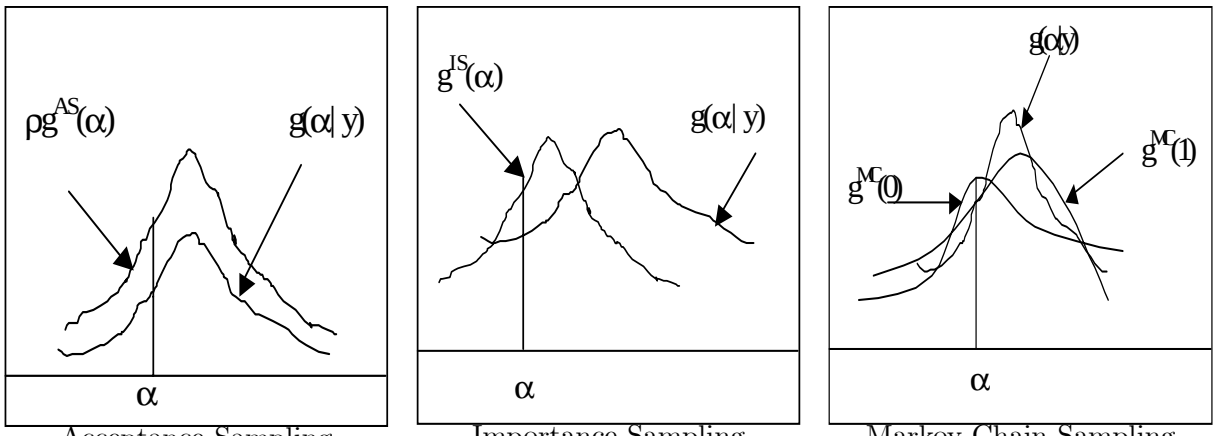
Let $g^{AS}(\alpha)$ be any function from which it is easy to simulate, defined for all $\alpha \in A$ for which $\breve{g}(\alpha|y) > 0$. Assume $\int_R g^{AS}(\alpha)d\alpha < \infty$ (not necessarily equal to 1) and that $IR^l = \frac{g(\alpha^l|y)}{g^{AS}(\alpha^l)} \leq \varrho < \infty, \ \ \forall \alpha \in A, \ l = 1, \ldots, L$. The left panel of figure 9.3 illustrates these assumptions. We want $\varrho g^{AS}(\alpha)$ uniformly above and approximately at the same distance from $g(\alpha|y)$ for every $\alpha \in A$. To generate an iid sequence from $g(\alpha|y)$, choose a $\varrho > 0$ and use the following:

Algorithm 9.5

1) *Draw* $\alpha^\dagger$ *from* $g^{AS}(\alpha)$ *and* U *from a* $\mathsf{U}(0,1)$.

2) *If* $\mathsf{U} > \frac{\breve{g}(\alpha^l|y)}{\varrho g^{AS}(\alpha^l)}$ *repeat 1); else set* $\alpha^l = \alpha^\dagger$.

3) *Repeat 1)-2) L times.*



Acceptance Sampling Importance Sampling Markov Chain Sampling

Figure 9.3: Posterior Simulators

To make algorithm 9.5 operative we need to select ϱ. Ideally $g^{AS}(\alpha) \propto \breve{g}(\alpha|y)$, so ϱ is a constant. In practice, ϱ will be varying across draws, hopefully not too much. Since the expected acceptance rate is $\frac{1}{\varrho}$, the optimal value of ϱ is $\sup_\alpha \frac{g(\alpha|y)}{g^{AS}(\alpha)}$ and one fiddles with ϱ until a 40-50% acceptance rate is achieved. Notice that algorithm 9.5 is self-monitoring: if ϱ is too large, it will reject frequently; if it is too small, it will accept all draws. Typical choices of $g^{AS}(\alpha)$ are t-distribution, split t-distribution for problems with normal errors and exponential or Beta-distribution for problems with binomial/ multinomial errors.

Example 9.29 *(Consumption function) Let* $c_t = GDP_t\alpha + e_t$, $e_t \sim \mathsf{N}(0, \sigma_e^2)$, σ_e^2 *fixed and let* $g(\alpha) \propto exp[-.5(\alpha-\bar{\alpha})'\bar{\Sigma}_\alpha^{-1}(\alpha-\bar{\alpha})]$ *if* $\alpha > 0$ *and 0 otherwise (positive marginal propensity to consume). The posterior kernel is* $exp[-.5(\alpha-\tilde{\alpha})'\tilde{\Sigma}_\alpha^{-1}(\alpha-\tilde{\alpha})]\mathcal{I}_{(\alpha>0)}$ *where* $\mathcal{I}(.)$ *is an indicator function,* $\tilde{\Sigma}_\alpha = (\bar{\Sigma}_\alpha^{-1} + \sigma_e^{-2}GDP'GDP)^{-1}$, $\tilde{\alpha} = \tilde{\Sigma}_\alpha(\bar{\Sigma}_\alpha^{-1}\bar{\alpha} + \sigma^{-2}GDP'GDP\alpha_{ols})$, *where* α_{ols} *is the OLS estimator. Suppose* $g^{AS}(\alpha)$ *is* $\mathsf{N}(\tilde{\alpha}, \tilde{\Sigma}_\alpha)$. *Then a draw is accepted if* $\alpha^\dagger > 0$ *and rejected if* $\alpha^\dagger \leq 0$.

Exercise 9.23 *Consider studying the duration dependence of recessions. Negative duration dependence occurs if the longer you have been in a particular state the higher is the probability to switch away from that state. Suppose we model the duration of recessions as Weibull with shape parameter* a_1 *and scale parameter* a_2 *and assume an appropriate prior for* α_1 *and* α_2 *(e.g. Beta or Uniform). Using US post-WWII GDP data draw samples from the posterior for* a_1 *and* a_2 *keeping only draws which produce negative duration. What is the mean of* a_2*? What is a 68% credible posterior interval? How would you verify the hypothesis that* $a_2 = 0$ *(i.e. no duration dependence)?*

If $g^{AS}(\alpha)$ is far away from $g(\alpha|y)$, sampling is time consuming since many draws will be discarded. The alternative is to keep all the draws but weight them appropriately.

Importance Sampling

Let $g^{IS}(\alpha)$ be an importance sampling density and $\breve{g}^{IS}(\alpha)$ be its kernel. Let $IR(\alpha) = \frac{g(\alpha|y)}{g^{IS}(\alpha)}$ be a weighting function with finite expected value. If $E(h(\alpha)|y)$ and $var(h(\alpha)|y)$ exist for a continuous $h(\alpha)$ and the support of $g^{IS}(\alpha)$ includes the support of $g(\alpha|y)$, then:

$$h_L \equiv \frac{\sum_{l=1}^{L} h(\alpha^l)IR(\alpha^l)}{\sum_l IR(\alpha^l)} \xrightarrow{P} E(h(\alpha)|y) \tag{9.37}$$

$$\sqrt{L}(h_L - E[h(\alpha)|y]) \xrightarrow{D} \mathsf{N}(0, \sigma^2) \tag{9.38}$$

$$\sigma_L^2 = \frac{L^{-1}\sum_{l=1}^{L}[h(\alpha^l) - E(h(\alpha)|y)]^2 IR(\alpha^l)^2}{[\sum_l IR(\alpha^l)]^2} \xrightarrow{P} \sigma^2 \tag{9.39}$$

An importance sampling density is pictured in the second panel of figure 9.3. Equation (9.37) implies that the acceptance sampling algorithm 9.5 can be simplified as follows:

Algorithm 9.6

1) *Draw* $\alpha^\dagger$ *from* $g^{IS}(\alpha)$*; weight* $h(\alpha^\dagger)$ *with* $\frac{g(\alpha^\dagger|y)}{g^{IS}(\alpha^\dagger)} = IR(\alpha^\dagger)$.

2) *Repeat step 1. L times and compute (9.37).*

Acceptance and importance sampling have similarities and differences. Importance sampling has a wider range of applicability since $IR(\alpha^l)$ can be computed using the kernels of

$g(\alpha|y)$ and $g^{IS}(\alpha)$. However, $g^{IS}(\alpha)$ must integrate to 1. Note that, if $\int g^{AS}(\alpha)d\alpha = 1$ and $g^{AS}(\alpha)$ satisfies $\frac{\breve{g}(\alpha^l|y)}{g^{AS}(\alpha^l)} \leq \varrho < \infty$, all l, it can be used as importance sampling.

To compute the marginal posterior for the subvector α_1, $g(\alpha_1|y) = \int\int g(\alpha_1, \alpha_2|y)d\alpha_2$, one could use $g(\alpha_1|y) = \int \frac{g(\alpha_1,\alpha_2|y)}{IR(\alpha_1|\alpha_2,y)} IR(\alpha_1|\alpha_2, y)d\alpha_2 \equiv E_{IR}\frac{g(\alpha_1,\alpha_2|y)}{IR(\alpha_1|\alpha_2 y)}$. Note also that if $IR(\alpha)$ is chosen so that $\frac{g(\alpha|y)}{IR(\alpha)g^{IS}(\alpha)}$ is constant, the algorithm works well. While a good importance density for interesting macroeconomic problems can sometimes be found (see example 3.2 in chapter 10), there are situations where $IR(\alpha)$ varies wildly, making this posterior simulator unusable. Since the properties of $IR(\alpha)$ are application dependent, careful experimentation is needed before the results obtained with importance sampling are to be trusted.

Finally note that, regardless of the variations in $IR(\alpha^l)$, both acceptance and importance sampling are difficult to use in hierarchical models or in structures where the dimension of α is large (e.g. in VARs). The methods described next can be used in both of these situations.

Example 9.30 *Continuing with example 9.29, let $g(\alpha)$ be defined for all $\alpha \in (-\infty, \infty)$. Then one could use a t-density with $(T-dim(\alpha))$ degrees of freedom as importance sampling. Since $IR(\alpha)$ is now the ratio of a normal to a t-density, it is bounded for all α. Hence, a t-density is a good importance density for this example.*

Exercise 9.24 *A simple model of returns states that $R_{it} = R_{Mt}\alpha_i + e_{it}$ where $i = 1, \ldots I, t = 1, \ldots T$ and R_{Mt} is a market portfolio (say, the return on the SP500 index). Suppose that, because of risk considerations, the prior for α is normal truncated outside the range (-2,2), i.e. $g(\alpha_i) = N(\bar{\alpha}, \bar{\sigma}_\alpha^2) * \mathcal{I}_{[-2,2]}$. Describe how to implement an importance sampling algorithm to construct $g(\alpha|R_i, R_M)$. How would you select a portfolio composed of assets whose returns are positively correlated with the market?*

Exercise 9.25 *Suppose $y_t = a_0 + \sum_{j=1}^{q} a_j y_{t-j} + e_t$ and suppose you are interested in $h(\alpha) = \sum_{j=1}^{q} \alpha_j$, a measure of the persistence of the process. Assume that $g(\sum_j \alpha_j) \sim U(0,1)$. Using EU data on CPI inflation from 1970 to 2000, draw a posterior sample for $h(\alpha)$ using both acceptance and importance sampling. What is the interquartile range for $h(\alpha)$ in the two cases? (Hint: make appropriate assumptions about e_t and its variance σ_e^2).*

9.5.3 Markov Chain Monte Carlo Methods

Monte Carlo Markov Chain (MCMC) methods are simulation techniques that generate a sample from some target distribution. The idea is to specify a transition kernel for a Markov chain such that starting from some initial value and iterating a number of times, we produce a limiting distribution which is the target distribution we need to sample from. The Metropolis-Hastings (MH) algorithm was the first MCMC method employed in the literature. Here the next value of the chain is generated from a proposal density and accepted or rejected according to the value of the target density at the candidate point relative to the target density at the current point. The Gibbs sampler, a method commonly used for a variety of economic problems, is a special case of the MH algorithm where draws for the subcomponents are made from a sequence of conditional distributions.

The generated sample can be used to summarize the target density using graphical methods and expectations of integrable functions can be estimated using appropriate averages of the functions. Under general conditions, the ergodicity of the Markov chain guarantees that this estimate is consistent and has a normal distribution as the length of the simulation goes to infinity. Note that, while with acceptance and importance sampling draws are *iid*, here the draws are correlated because of the Markov nature of the process. Therefore averages should be computed from approximately independent elements of the sequences or the asymptotic covariance matrix appropriately modified.

MCMC methods can be applied directly to the kernel of the target density (that is, no knowledge of $f(y)$ is needed) and this makes them particularly useful for Bayesian analysis. However, as we will see in the next chapters, MCMC methods can also be used as classical devices to explore intractable likelihoods or to find the maximum of nasty functions using a "data augmentation" technique.

To see how the method works take as the limiting distribution $\mu(\alpha) \equiv \breve{g}(\alpha|y)$. Then we need a transition $P(A_s, \alpha)$, where $A_s \subseteq A$, which converges as the number of iterations goes to infinity to $\mu(\alpha)$, starting from any α_0. Suppose $P(d\alpha', \alpha) = p(\alpha', \alpha)d\alpha' + p_1(\alpha)p_2(d\alpha')$ for some p, where $\alpha' \in A_s$, $p(\alpha, \alpha) = 0$, $p_2(d\alpha')$ has a point mass at α (i.e. it equals one if $\alpha \in d\alpha'$) and $p_1(\alpha) = 1 - \int p(\alpha, \alpha')d\alpha'$ is the probability that the chain remains at α. Suppose $\mu(\alpha)p(\alpha, \alpha') = \mu(\alpha')p(\alpha', \alpha)$ (this condition is called reversibility). Then

$$\int P(A_s, \alpha)\mu(\alpha)d\alpha = \int_{A_s} \mu(\alpha')d\alpha' \tag{9.40}$$

Exercise 9.26 *Show that (9.40) holds (Hint: use the reversibility condition and the fact that* $\int p(\alpha', \alpha)\mu(\alpha')d\alpha = (1 - p_1(\alpha'))\mu(\alpha')$*.)*

The left hand side of (9.40) is the unconditional probability of going from α to $\alpha' \in A_s$ where α is generated from $\mu(\alpha)$. The right hand side is the unconditional probability of being in α' where α' is generated by $\mu(\alpha')$. Condition (9.40) defines an invariant distribution μ for $P(A_s, \alpha)$. Therefore, if $\mu(\alpha)$ is unique, $P(A_s, \alpha)$ is chosen as above and iterated L times, the result will be the target distribution (see last panel of figure 9.3 for the first two steps in the iterations). To show the details of the argument we need a few definitions.

Definition 9.5 *A Markov chain is a collection of random variables* $\alpha_t, t = 1, \ldots T$*. The transition matrix of a Markov chain is* $P(A, \alpha^\ddagger) = pr(\alpha' \in A|\alpha = \alpha^\ddagger) = \int_A \mathcal{K}(d\alpha', \alpha^\ddagger)$ *where* $\mathcal{K}$ *is the kernel of the chain,* $\mathcal{K}(., \alpha^\ddagger)$ *is a probability measure for all* $\alpha^\ddagger$*, and* $\mathcal{K}(A, .)$ *is measurable for all* A*. The* L*-step transition matrix is* $P^L(A, \alpha^\ddagger) = pr(\alpha^L \in A|\alpha = \alpha^\ddagger) = \int \mathcal{K}(d\alpha', \alpha^\ddagger)\mathcal{K}^{L-1}(A^{L-1}, \alpha')$ *with* $\mathcal{K}^1(d\alpha', \alpha^\ddagger) = \mathcal{K}(d\alpha', \alpha^\ddagger)$*.*

Definition 9.6 *Let* $A_1 = \{\alpha \in A, pr(\alpha) > 0\}$*. The kernel of a Markov chain is irreducible if there exists a* $L \geq 1$ *such that* $\mathcal{K}^L(A_1, \alpha^\ddagger) > 0$*, for all* $\alpha^\ddagger \in A$*.*

Definition 9.7 *An irriducible Markov chain is aperiodic if for all* $A_2 \in A$*,* $P(A_2, \alpha^\ddagger) > 0$*,* $\forall \alpha^\ddagger \in A$*.*

Definition 9.8 *A Markov chain is Harris recurrent if there exists a measure P such that the kernel of the chain is irreducible and, for every A_3 with $p(A_3) > 0$, $P(A_3, \alpha^{\ddagger}) = 1$.*

Definition 9.9 *A function $\mu(\alpha)$ is an invariant density for the kernel of the Markov chain if $\mu(A_4) \equiv \int_{A_4} \mu(\alpha)d\alpha = \int \mathcal{K}(A_4, \alpha^{\ddagger})\mu(d\alpha^{\ddagger})$, for all measurable $A_4 \subseteq A$.*

The meaning of the irriducibility condition is shown in figure 9.4. The sequences in the first box stays within a particular region. Therefore, there is a part of the space which has zero probability of being visited starting either from A or B. This does not happen in the second box. The aperiodicity condition implies that all states can be visited with positive probability from any initial state. That is, we don't want the chain to cycle through a finite number of sets. Finally, a Harris recurrent chain visits A_3 with probability 1.

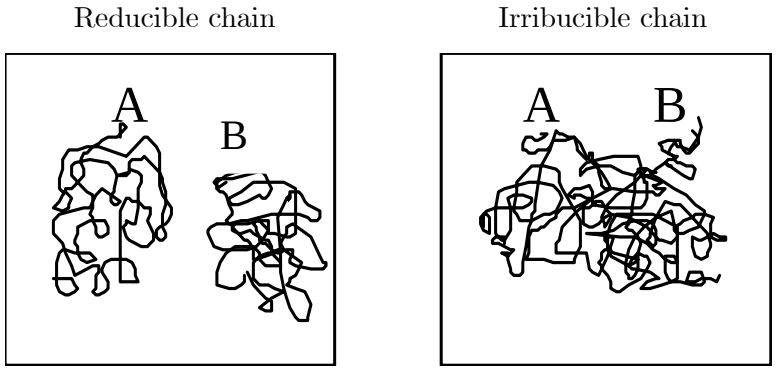


Figure 9.4: MCMC draws

With these definitions we can present the two main results which justify the use of MCMC methods to draw sequences from unknown posterior distributions.

Result 9.3 *(Tierney) If a Markov chain is Harris recurrent and has a proper ergodic $\mu(\alpha)$, than $\mu(\alpha)$ is the unique invariant distribution of the Markov chain.*

Result 9.4 *(Tierney) If a Markov chain with invariant distribution $\mu(\alpha)$ is Harris recurrent and aperiodic then, for all $\alpha_0 \in A$ and all A_0,*

1. *$||P^L(A_0, \alpha_0) - \mu(\alpha)|| \to 0$ as $L \to \infty$ where $||.||$ is the total variation distance.*
2. *For all $h(\alpha)$ absolutely integrable with respect to $\mu(\alpha)$: $\lim_{L\to\infty} \frac{1}{L} \sum_{l=1}^{L} h(\alpha^l) \xrightarrow{a.s.} \int h(\alpha)\mu(\alpha)d\alpha$.*

Part 1 of result 9.4 tells us that, as $L \to \infty$, draws from $P^L(\alpha_0, A_0)$ are draws from the invariant distribution, regardless of the initial value α_0. Part 2 indicates that averages of functions evaluated at sample values converge to their expected values calculated using the invariant distribution. Sufficient conditions which insures that the chain is Harris recurrent and aperiodic are given below for each posterior simulators.

Gibbs sampler

Given that the object is to find a transition density that has the joint posterior as its invariant distribution, the Gibbs sampler partitions the vector of unknowns into blocks and the transition density is defined by the product of conditional densities. The next item in the chain is obtained by successively sampling from the densities of each block, given the most recent values of the conditioning parameters. The main value of the algorithm lays in the fact that conditional densities are typically easy to compute and cheap to sample from.

To see exactly what the algorithm involves, partition α as $\alpha = (\alpha_1, \alpha_2, ..., \alpha_k)$. Suppose $g(\alpha_i|\alpha_{i'}, y\ i' \neq i)$ are available (e.g. choose the partition so this is the case). Then:

Algorithm 9.7

1) *Choose initial values* $(\alpha_1^{(o)}, \alpha_2^{(o)}, ..., \alpha_k^{(o)})$ *from an approximate* $g(\alpha|y)$*, e.g. a normal approximation or the output of another (simpler) simulator.*

2) *For* $l = 1, 2, \ldots,$ *draw* α_1^l *from* $g(\alpha_1|\alpha_2^{l-1}, ..., \alpha_k^{l-1}, y)$*,* α_2^l *from* $g(\alpha_2|\alpha_1^l, ..., \alpha_k^{l-1}, y), \ldots, \alpha_k^l$ *from* $g(\alpha_k|\alpha_1^l, ..., \alpha_{k-1}^l, y)$*.*

3) *Repeat step 2) L times.*

The process of drawing in step 2 defines a transition from α^{l-1} to α^l. The algorithm therefore produces a sequence which is the realization of a Markov chain with transition

$$P(\alpha^l, \alpha^{l-1}) = \prod_{i=1}^{k} g(\alpha_i^l|\alpha_{i'}^{l-1}(i' > i), \alpha_{i'}^l(i' < i), y) \tag{9.41}$$

By result 9.4, if $P(\alpha^l, \alpha^{l-1})$ is Harris recurrent and aperiodic, the sample $\alpha^L = (\alpha_1^L, \alpha_2^L, ..., \alpha_k^L)$, L is large, is a draw from the joint posterior $g(\alpha|y)$. Furthermore, $\alpha_i^L, i = 1, \ldots k$ is a draw from the marginal $g(\alpha_i|y)$.

The Gibbs sampler works well when the components are independent. Therefore highly correlated components (e.g. the parameters of an AR process) should be grouped together in blocks. Tractable conditional structures from intractable likelihood functions can be derived at times using a data-augmentation technique. An example of this technique is given below while applications to factor and Markov switching models are in chapter 11.

What kind of conditions insure that the transition kernel (9.41) converges to the posterior $g(\alpha|y)$? A sufficient condition is the following: if for every $\alpha_0 \in A$ and every $A_1 \subset A$ with $pr(\alpha \in A_1|y) > 0$, $P(\alpha^l \in A_1|\alpha^{l-1}, y) > 0$, where P is the transition induced by (9.41), then the Gibbs transition kernel is ergodic and its unique invariant distribution is $g(\alpha|y)$.

The condition $P(\alpha^l \in A_1|\alpha^{l-1}, y) > 0$ is simple and easy to check. In fact, it requires that all the cells of the chain can be visited with positive probability starting from any α^{l-1}. All the applications we discuss in chapters 10 and 11 satisfy this mild condition.

There are a few implementation issues worth discussing. The first is how to draw uncorrelated samples for α from $g(\alpha|y)$. There are two alternatives. The first one produces

one sample (of dimension $J * L$) after an initial sequences of $\bar{L}$ observations is discarded. Then one uses only elements $(L, 2L, \ldots, J * L)$ to eliminate the correlation existing among the draws. The second produces J samples each of length $L + \bar{L}$ and the last observation in each sample is used for inference. If $\bar{L}$ is chosen appropriately, the two approaches are equivalent. The second important issue has to do with the magnitude of $\bar{L}$, the length of the burn-out period. There are many ways to check how long $\bar{L}$ should be to insure that the algorithm has converged. Here we describe three: the first two are appropriate when draws for α are made from one large sample. The last is applicable when J samples of L observations are used.

One way to check for convergence is to choose two points, say $\bar{L}_1 < \bar{L}_2$, and compute distributions/moments of α after these points. If the distributions/moments are stable, then the algorithm has converged at $\bar{L}_1$. Taking this approach one step further, one can compute recursive means of α^l over l and graphically check if they settle after an initial period (CUMSUM statistic). Alternatively, one could fix $\bar{L}$ and compute distributions/moments using J_1 and J_2 sampled values, $J_2 >> J_1$. If convergence is achieved, the distributions/moments computed with J_1 observations should be similar to those computed using J_2 observations. A variant of this approach is the following: let $h(\alpha)$ be a continuous function of α. Then, given $\bar{L}$, one could also split the simulation sample into two pieces $(J_1 * L)$ and $(J_2 * L)$, $J = J_1 + J_2$ and compute $h_1 = \frac{1}{J_1} \sum_l h(\alpha^l)$; $h_2 = \frac{1}{J_2} \sum_l h(\alpha^l)$; $\sigma_1^2 = \frac{1}{J_1} \sum_l [h(\alpha^l) - h_1]^2$; $\sigma_2^2 = \frac{1}{J_2} \sum_l [h(\alpha^l) - h_2]^2$. Convergence obtains if observations in the two samples have the same distribution, i.e. $\frac{h_1 - h_2}{(\sigma_1^2 + \sigma_2^2)^{0.5}} \xrightarrow{D} \mathsf{N}(0, 1)$ as $J \to \infty$. Both $\bar{L}$ and J are application dependent. For simple problems $\bar{L} \approx 50$ and $J \approx 200$ should suffice. For more complicated problems (for example, VARs or panel VARs), $\bar{L} \approx 100$ and $J \approx 300 - 500$ should be selected.

The third approach examines whether the variance within iteration is approximately the same as the variance across iterations. Failure to converge is indicated by the former being significantly smaller than the latter. That is, compute $\Sigma_B = \frac{L}{J-1} \sum_j (\bar{h}_{\cdot j} - \bar{h}_{\cdot\cdot})^2$ where $\bar{h}_{\cdot j} = \frac{1}{L} \sum_i h(\alpha_{ij})$; $\bar{h}_{\cdot\cdot} = \frac{1}{J} \sum_j h(\alpha_{\cdot j})$; and $\Sigma_W = \frac{1}{J} \sum_j (\frac{1}{L-1} \sum_i (h(\alpha_{ij}) - \bar{h}_{\cdot j}))^2$. Then $\sqrt{\frac{\frac{L-1}{L}\Sigma_W + \frac{1}{L}\Sigma_B}{\Sigma_W}} \to 1$ as $L \to \infty$. Hence, if $\Sigma_B \approx \Sigma_W$, convergence is achieved.

Example 9.31 *We examine convergence of Gibbs sampler estimates in a linear regression model where the log of output is regressed on a number of lags of the log of money. This could be of some interest, for example, in studying money neutrality in the short or in the long run. Data for the US from 1973:1 to 1993:12 are used in the exercise. We have run fifty replications of the Gibbs sampler using 150, 300, 500 draws for a model with one intercept and two lags of the log of money. The adjusted ratio of Σ_W to Σ_B was respectively 1.01, 1.003, 1.001 indicating that convergence was achieved after 150 draws. For each of the replications with 500 draws, we split the sample in two with 300 observations in the first part and 200 in the second part and computed the normal test. Out of 100 replications, we rejected the null of convergence in just one case.*

Inference with the output of the Gibbs sampler presents no difficulty. As suggested by

result 9.4, $E(h(\alpha|y)) = \frac{1}{J}\sum_j h(\alpha^{jL})$ where the notation α^{jL} indicates the $j*L$-th draw after $\bar{L}$ iterations are performed. The variance of $h(\alpha)$ can be computed using the spectral density at frequency zero, i.e. $E(h(\alpha|y)h(\alpha|y)') = \sum_{-J(\tau)}^{J(\tau)} \mathcal{K}(\tau)ACF_h(\tau)$, where $ACF_h(\tau)$ is the autocovariance of $h(\alpha)$ for draws separated by τ periods and $J(\tau)$ is the maximum number of covariances considered. Note that this measure takes into account the possibility that the selected draws are not spaced enough to make them independent of each other. The marginal density for α_i can be estimated using kernel methods directly from the sequence $(\alpha_i^1, \dots \alpha_i^J)$ or using the fact that $g(\alpha_i|y) = \frac{1}{J}\sum_{j=1}^{J} g(\alpha_i|y, \alpha_{i'}^j, i' \neq i)$.

Predictive inference is also easy. For example, $f(y_{t+\tau}|y_t) = \int f(y_{t+\tau}|y_t, \alpha)g(\alpha|y_t)d\alpha$ can be easily simulated using Gibbs sampler draws for α and the model specification $f(y_{t+\tau}|y_t, \alpha)$, averaging simulated values of $y_{t+\tau}$ over α draws. Finally, tests of model adequacy can also be implemented using the output of the Gibbs sampler. Recall that the Bayes factor is the ratio of the predictive density of two models. Hence, it can be numerically calculated drawing α from $g(\alpha)$, constructing $f(y|\alpha)$ for each draw, and averaging over α's for each of the two models.

We illustrate the properties of the Gibbs sampler in a simple example next.

Example 9.32 *Suppose that (x, y) is binomial with density $f(x, y) \propto \frac{n!}{x!(n-x)!}y^{x+\alpha_0-1}(1-y)^{n-x+\alpha_1-1}$, $x = 0, 1, \dots n$, $0 \leq y \leq 1$ and suppose we are interested in the marginal $f(x)$. Direct integration produces $f(x) \propto \frac{n!}{x!(n-x)!}\frac{\Gamma(\alpha_0+\alpha_1)}{\Gamma(\alpha_0)\Gamma(\alpha_1)}\frac{\Gamma(x+\alpha_0)\Gamma(n-x+\alpha_1)}{\Gamma(\alpha_0+\alpha_1+n)}$ which is the Beta-Binomial distribution. It is also easy to calculate the conditional distributions $f(x|y)$ and $f(y|x)$. The first one is binomial with parameters (n, y), the second is Beta with parameters $(x+\alpha_0, n-x+\alpha_1)$. Figure 9.5 presents the histogram generated from the true $f(x)$ when $\alpha_0 = 2$, $\alpha_1 = 4$ and from the marginal computed via the Gibbs sampler with $J = 500$, $L = 100$, and $\bar{L} = 20$. It is remarkable how close the two distributions are, even for a small $\bar{L}$.*

In the next example we use the Gibbs sampler to obtain the posterior distribution of the parameters of a seemingly unrelated regression (SUR) model - it is a good idea to store these derivations as we will extensively use them in chapter 10.

Example 9.33 *(Seemingly unrelated regression) Let $y_{it} = x_{it}'\alpha_i + e_{it}$ and $e_t = (e_{1t}', \dots, e_{Mt}')' \sim \mathsf{N}(0, \Sigma_e)$, where $i = 1, \dots, m$, $t = 1, \dots, T$ and α_i is a $k \times 1$ vector. Stacking the observations for each i we have $y_t = x_t\alpha + e_t$ where $y_t = (y_{it}', \dots, y_{mt}')', x_t = \text{diag}(x_{it}', \dots, x_{mt}')$, $\alpha = (\alpha_1', \dots \alpha_m')'$ is an $mk \times 1$ vector. Suppose that $g(\alpha, \Sigma_e^{-1}) = g(\alpha)g(\Sigma_e^{-1})$. Then:*

$$\breve{g}(\alpha, \Sigma_e^{-1}|y) = g(\alpha)g(\Sigma_e^{-1})|\Sigma_e^{-1}|^{0.5T}\exp\{-0.5\sum_t (y_t - x_t\alpha)'\Sigma_e^{-1}(y_t - x_t\alpha)\} \tag{9.42}$$

The target density that needs to be simulated is $\frac{\breve{g}(\alpha, \Sigma_e^{-1}|y)}{\int\int \breve{g}(\alpha, \Sigma_e^{-1}|y)d\alpha d\Sigma_e}$. Assume a conjugate prior for α and Σ_e^{-1} of the Normal-Wishart form. Then $g(\alpha|Y, \Sigma_e^{-1}) \sim \mathsf{N}(\tilde{\alpha}, \tilde{\Sigma}_\alpha)$ and $g(\Sigma_e^{-1}|\alpha, Y) \sim \mathsf{W}(T+\bar{\nu}, \tilde{\Sigma})$ where $\tilde{\alpha} = \tilde{\Sigma}_\alpha(\bar{\Sigma}_\alpha^{-1}\bar{\alpha} + \sum_t x_t'\Sigma_e^{-1}y_t)$; $\tilde{\Sigma}_\alpha = (\bar{\Sigma}_\alpha^{-1} + \sum_t x_t'\Sigma_e^{-1}x_t)^{-1}$ and $\tilde{\Sigma} = (\bar{\Sigma}^{-1} + \sum_t (y_t - x_t\alpha_{ols})(y - x_t\alpha_{ols})')^{-1}$, where $(\bar{\alpha}, \bar{\Sigma}_\alpha)$ are the prior mean and variance of α, Σ is the scale matrix of the prior for Σ_e^{-1}, $\bar{\nu}$ the prior degrees of freedom and

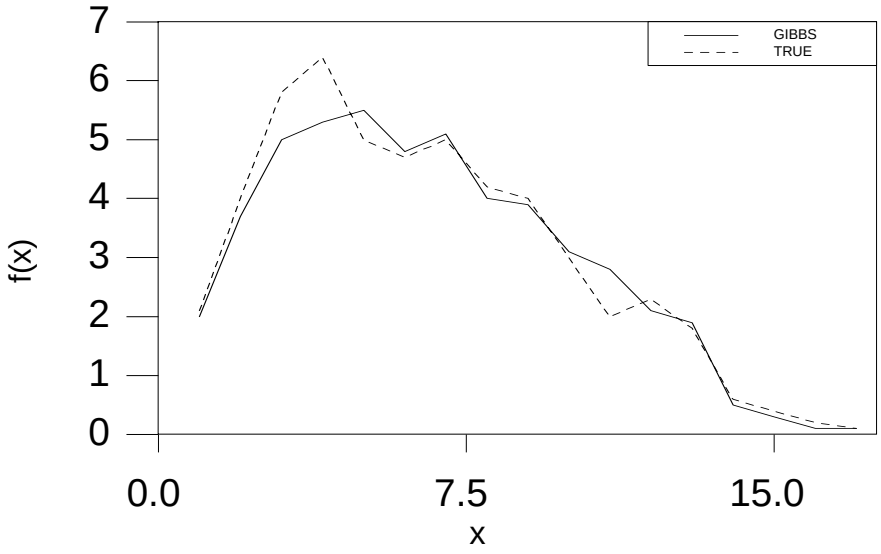


Figure 9.5: True and Gibbs sampling distributions

α_{ols} is the OLS estimator of α. If we treat α and Σ_e as two Gibbs sampler blocks, simulations of the two conditional posteriors asymptotically yield a sample such that $\alpha^{jL} \sim g(\alpha|y)$; $\Sigma_e^{-1(jL)} \sim g(\Sigma_e^{-1}|y)$ and $(\alpha^{jL}, \Sigma_e^{-1(jL)}) \sim g(\alpha, \Sigma_e^{-1}|y)$.

Exercise 9.27 *Suppose in example 9.33 that $m = 1$, i.e. $y_t = x_t\alpha + e_t$, $e_t \sim \mathsf{N}(0, \sigma_e^2)$. i) Assume a non-informative prior for σ_e^{-2} and that $\alpha \sim \mathsf{N}(\bar{\alpha}, \bar{\Sigma}_\alpha)$. Calculate the conditional posterior for α and σ_e^{-2}. Describe how to employ the Gibbs sampler in this situation. ii) Suppose that $e_t \sim \mathsf{N}(0, \sigma_e^2 * x'x)$ and assume that the priors are $\bar{s}^2\sigma^{-2} \sim \chi^2(\bar{\nu})$ and that $\alpha \sim \mathsf{N}(\bar{\alpha}, \bar{\Sigma}_\alpha)$. Find the conditional posterior for α and σ_e^2 and show how to implement the Gibbs sampler in this case. (Hint: Transform the model to get rid of heteroschedasticity).*

Example 9.34 *(Variance models) Consider the model $y_{it} = \alpha_i + e_{it}$, $i = 1, \ldots n$, $t = 1, \ldots T$, where $\alpha_i \sim \mathsf{N}(\bar{\alpha}, \bar{\sigma}_\alpha^2)$, $e_{it} \sim \mathsf{N}(0, \sigma_e^2)$. Let $\alpha = [\alpha_1, \ldots \alpha_n]'$, $y = [y_{11}, \ldots y_{nT}]'$ and assume that $\bar{\alpha} \sim \mathsf{N}(\bar{\alpha}_0, \sigma_0^2)$, $\bar{\sigma}_\alpha^{-2} \sim \mathsf{G}(a_1^\alpha, a_2^\alpha)$, $\sigma_e^{-2} \sim \mathsf{G}(a_1^e, a_2^e)$ where $(\sigma_0^2, \bar{\alpha}_0, a_1^\alpha, a_2^\alpha, a_1^e, a_2^e)$ are known. Then the conditional posteriors are $(\bar{\sigma}_\alpha^{-2}|y, \bar{\alpha}, \alpha, \sigma_e^{-2}) \sim \mathsf{G}(a_1^\alpha + 0.5n, a_2^\alpha + 0.5 \sum_i(\alpha_i - \bar{\alpha})^2))$; $(\sigma_e^{-2}|y, \bar{\alpha}, \alpha, \bar{\sigma}_\alpha^{-2}) \sim \mathsf{G}(a_1^e + 0.5nT, a_2^e + 0.5 \sum_i \sum_t (y_{it} - \alpha_i)^2)$; $(\bar{\alpha}|\bar{\sigma}_\alpha^2, y, \alpha, \sigma_e^2) \sim \mathsf{N}(\frac{\bar{\sigma}_\alpha^2\bar{\alpha}_0 + \sigma_0^2 \sum_i \alpha_i}{\bar{\sigma}_\alpha^2 + n\sigma_0^2}, \frac{\bar{\sigma}_\alpha^2\sigma_0^2}{\bar{\sigma}_\alpha^2 + n\sigma_0^2})$; and $(\alpha|\bar{\sigma}_\alpha^2, y, \bar{\alpha}, \sigma_e^2) \sim \mathsf{N}(\frac{T\sigma_\alpha^2}{T\bar{\sigma}_\alpha^2 + \sigma_e^2}\bar{y} + \frac{\sigma_e^2}{T\bar{\sigma}_\alpha^2 + \sigma_e^2}\bar{\alpha}\mathbf{1}, \frac{\bar{\sigma}_\alpha^2\sigma_e^2}{T\bar{\sigma}_\alpha^2 + \sigma_e^2} I)$ where $\bar{y} = (\bar{y}_1, \ldots \bar{y}_n)'$, $\bar{y}_i = \frac{1}{T} \sum_t y_{it}$, $\mathbf{1}$ is a $n \times 1$ vector of ones and I the identity matrix.*

Exercise 9.28 *Let $y_{it} \sim \mathsf{N}(\alpha_i, \sigma_i^2)$, $i = 1, \ldots, n$, $t = 1, \ldots T_i$; where $\alpha_i \sim \mathsf{N}(\bar{\alpha}, \bar{\sigma}_\alpha^2)$; $\sigma_i^{-2} \sim \mathsf{G}(a_1^i, a_1^i)$ $\bar{\sigma}_\alpha^{-2} \sim \mathsf{G}(a_1^\alpha, a_2^\alpha)$, $\bar{\alpha} \sim \mathsf{N}(\bar{\alpha}_0, \sigma_0^2)$ and $(a_1^i, a_2^i, a_1^\alpha, a_2^\alpha, \bar{\alpha}_0, \sigma_0^2)$ are known. Let $\bar{y}_i = \frac{1}{T_i} \sum_t y_{it}$, $s_i^2 = \frac{1}{T_i - 1} \sum_t (y_{it} - \bar{y}_i)^2$, $\alpha = (\alpha_1, \ldots, \alpha_n)'$, $Y = (\bar{y}_1, \ldots \bar{y}_n, s_1^2, \ldots s_n^2)'$,*

$\sigma^2 = (\sigma_1^2, \ldots \sigma_n^2)'$.
(i) Show that $(\alpha_i|Y,\ \sigma^2,\ \bar{\alpha},\ \bar{\sigma}_\alpha) \sim \mathsf{N}(\frac{T_i\bar{y}_i\bar{\sigma}_\alpha^2+\bar{\alpha}\sigma_i^2}{T_i\bar{\sigma}_\alpha^2+\sigma_i^2}; \frac{\sigma_i^2\bar{\sigma}_\alpha^2}{T_i\bar{\sigma}_\alpha^2+\sigma_i^2})$ *(Note* $cov(\alpha_i, \alpha_j) = 0$*);* $(\sigma_i^{-2}|Y, \alpha, \bar{\alpha},$ $\sigma_\alpha^2) = \prod_i \mathsf{G}(a_1^i + 0.5T_i, a_2^i + \sum_{t=1}^{T_i}(y_{it} - \alpha_i)^2)$*;* $(\bar{\alpha}|Y, \alpha, \bar{\sigma}_\alpha^2, \sigma^2) \sim \mathsf{N}(\frac{\bar{\sigma}_\alpha^2\bar{\alpha}_0+\sigma_0^2\sum_i \alpha_i}{\bar{\sigma}_\alpha^2+n\sigma_o^2}; \frac{\sigma_0^2\bar{\sigma}_\alpha^2}{\bar{\sigma}_\alpha^2+n\sigma_0^2})$*;* $(\bar{\sigma}_\alpha^{-2}|Y, \alpha, \bar{\alpha}, \sigma_i^2) = \mathsf{G}[a_1^\alpha + 0.5n, a_2^\alpha + 0.5\sum_{i=1}^n(\alpha_i - \bar{\alpha})^2]$
(ii) Assume $\bar{\alpha}_0 = 0; \sigma_0^2 = 1000; a_1^i = 0.5, a_2^i = 1,\ \forall i, a_1^\alpha = a_2^\alpha = 0, n = 3$ *and suppose you have the following data* $T_i = (6, 8, 5); \bar{y}_i = (0.31, 2.03, 6.39); s_i^2 = (0.23, 2.47, 8.78)$*. Draw posterior samples for* α*. Produce the posterior of* $\alpha_2 - \alpha_1$ *and* $\alpha_3 - \alpha_1$*.*
(iii) Suppose a fourth unit is added to the sample with $T_4 = 2, \bar{y}_4 = 5.67, \sigma_4^2 = 4.65$*. Construct a time series of* 5 *observations for this new unit.*

The Gibbs sampler is very useful to evaluate likelihoods with latent variables.

Example 9.35 *(Latent variables) Consider the problem of modeling monthly purchases of certain non-durable goods. We have a sample with many consumers but only a fraction of them have acquired the good under consideration (say, tomatoes). We assume that agents purchase tomatoes on the basis of individual characteristics. Suppose we measure purchases in kilos and write the following censored regression model:* $z_i = x_i'\alpha + e_i,\ e_i\ iid \sim \mathsf{N}(0, \sigma_e^2)$*, and* $y_i = \max(0, z_i)$*. Here* z_i *is a latent variable. Given* n *consumers,* n_1 *of which buy tomatoes, the likelihood for* (α, σ^2) *is*

$$\mathcal{L}(\alpha, \sigma_e^2|y) = \prod_{i\in(n-n_1)}(1 - \mathsf{SN}(\frac{x_i'\alpha}{\sigma_e})) \prod_{i\in n_1} \sigma_e^{-2}\exp\{-0.5\sigma_e^{-2}(y_i - x_i'\alpha)^2\} \tag{9.43}$$

where SN *is a standard normal distribution. This function is difficult to manipulate. However, if we treat* z_i *as a latent variable and use the model for* z_i *to artificially augment the data space (we have called this approach data-augmentation technique), then the posterior distribution can be easily sampled from the conditionals of* (α, σ_e^2, z)*. If* $g(\alpha) \sim \mathsf{N}(\bar{\alpha}, \bar{\Sigma}_\alpha); \sigma_e^{-2} \sim \mathsf{G}(a_1, a_2)$ *then* $(\alpha|\bar{\sigma}_e^2, z, y) \sim \mathsf{N}(\tilde{\alpha}, \tilde{\Sigma}_\alpha)$ *and* $(\sigma_e^{-2}|\alpha, z, y) \sim \mathsf{G}(a_1+0.5n), a_2+ 0.5(y - x'\alpha_{ols})'(y - x'\alpha_{ols}))$*. Furthermore since* e_i *are iid,* $g(z|y, \alpha, \sigma_e^2) = \prod_{i\in(n-n_1)}$ $g(z_i|y, \alpha, \sigma_e^2) = \prod_{i\in(n-n_1)} \mathcal{I}_{(-\infty,0]}\mathsf{N}(x_i'\alpha, \sigma_e^2)$ *where* $\mathcal{I}_{(-\infty,0]}$ *truncates the normal distribution outside the support* $(-\infty, 0]$*. This simplification is possible because* (α, σ_e^2) *depend on* z_i *only through* y_i *(see Tanner and Wong (1987)).*

In applied work missing data causes headaches. However, if we treat them as latent variables, the Gibbs sampler can be used to reconstruct them.

Exercise 9.29 *(Missing data) Suppose we have missing data from a time series* y_t*. Let* y_t^M *be the missing data and* y_t^A *the data available and let* $y_t = [y_t^M, y_t^A]' = x_t\alpha + e_t$ *where* $e_t \sim \mathsf{N}(0, \Sigma_e)$*. Here* x_t *is a vector of observable variables;* $\Sigma_e = \begin{matrix}\sigma_1^2+\sigma_2^2 & \sigma_2^2 \\ \sigma_2^2 & \sigma_1^2+\sigma_2^2\end{matrix}$*. Assume a normal prior for* α *and a non-informative prior for* (σ_1^2, σ_2^2)*. Show that* $(y_t^M|y_t^A, \alpha, \sigma_1^2, \sigma_2^2)$ *is normal. Show the moments of the distribution. Describe how to use the Gibbs sampler to draw missing data. Explain why treating* y_t^M *as a vector of unknown parameters makes the posterior tractable.*

Metropolis-Hastings algorithm

The Metropolis-Hastings (MH) algorithm is a general simulation procedure which allows to sample from intractable distributions. Two are the typical applications of this algorithm: to sample blocks within a Gibbs sampler which have truncated distributions or distributions where only a part of the kernel is tractable; to sample in problems where the block structure needed to implement the Gibbs sampler is not available.

The MH algorithm works are follows: given that the latest value of α is α^{l-1}, the next value of the sequence is generated drawing from a candidate density $\mathsf{P}(\alpha^{\dagger}, \alpha^{l-1})$. The draw is accepted with a probability which depends on the ratio of values of $\breve{g}(\alpha^{\dagger}|y) \times \mathsf{P}(\alpha^{\dagger}, \alpha^{l-1})$ to $\breve{g}(\alpha^{l-1}|y) \times \mathsf{P}(\alpha^{l-1}, \alpha^{\dagger})$. If a candidate is rejected, $\alpha^{l} = \alpha^{l-1}$. Hence, starting from some $\alpha^{0} \in A$ and an arbitrary transition $\mathsf{P}(\alpha^{\dagger}, \alpha^{l-1})$ where $\alpha^{\dagger} \in A$, one proceeds as follows. For $l = 1, 2, \ldots L$

Algorithm 9.8

1) *Draw* $\alpha^{\dagger}$ *from* $\mathsf{P}(\alpha^{\dagger}, \alpha^{l-1})$ *and* U *from* $\mathsf{U}(0,1)$.

2) *If* $\mathsf{U} < \mathsf{E}(\alpha^{l-1}, \alpha^{\dagger}) = [\frac{\breve{g}(\alpha^{\dagger}|y)\mathsf{P}(\alpha^{\dagger}, \alpha^{l-1})}{\breve{g}(\alpha^{l-1}|y)\mathsf{P}(\alpha^{l-1}, \alpha^{\dagger})}]$ *set* $\alpha^{\ell} = \alpha^{\dagger}$, *else set* $\alpha^{\ell} = \alpha^{\ell-1}$.

The recursions produced by algorithm 9.8 define a Markov chain with a mixture of continuous and discrete transitions:

$$\begin{aligned} P(\alpha^{l-1},\ \alpha^{l}) &= \mathsf{P}(\alpha^{l}, \alpha^{l-1})\mathsf{E}(\alpha^{l-1}, \alpha^{l}) \text{ if } \alpha^{l} \neq \alpha^{l-1} \\ &= 1 - \int_{A} \mathsf{P}(\alpha, \alpha^{l-1})\mathsf{E}(\alpha^{l-1}, \alpha) d\alpha \ \ if\ \alpha^{l} = \alpha^{l-1} \end{aligned} \tag{9.44}$$

Note that if $\mathsf{P}(\alpha^{l-1}, \alpha^{\dagger}) = \mathsf{P}(\alpha^{\dagger}, \alpha^{l-1})$, the acceptance rate is independent of P and we have the Metropolis version of the algorithm.

The logic of algorithm 9.8 is simple. Suppose that at a particular draw $\breve{g}(\alpha^{\dagger}|Y)\mathsf{P}(\alpha^{\dagger}, \alpha^{l-1})$ $\leq \breve{g}(\alpha^{l-1}|Y)\mathsf{P}(\alpha^{l-1}, \alpha^{\dagger})$. Then, loosely speaking, the process moves too rarely from $\alpha^{\dagger}$ to α^{l-1} and too often from α^{l-1} to $\alpha^{\dagger}$. To counteract this effect we introduce a probability $\mathsf{E}(\alpha^{\dagger}, \alpha^{l-1})$ that the move is made so that the transition from α^{l-1} to $\alpha^{\dagger}$ is $\mathsf{P}(\alpha^{l-1}, \alpha^{\dagger})\mathsf{E}(\alpha^{l-1}, \alpha^{\dagger})$. Since the process does not move often enough from α^{l-1} to $\alpha^{\dagger}$ we set $\mathsf{E}(\alpha^{\dagger}, \alpha^{l-1}) = 1$. Using the reversibility conditions at equality, we have $\mathsf{E}(\alpha^{l-1}, \alpha^{\dagger}) = \frac{\breve{g}(\alpha^{\dagger}|y)\mathsf{P}(\alpha^{\dagger}, \alpha^{l-1})}{\breve{g}(\alpha^{l-1}|y)\mathsf{P}(\alpha^{l-1}, \alpha^{\dagger})}$. That is, $\mathsf{E}(\alpha^{l-1}, \alpha^{\dagger})$ is insures that the reversibility condition is satisfied.

Further intuition can be gained if $\mathsf{P}(\alpha^{l-1}, \alpha^{\dagger}) = \mathsf{P}(\alpha^{\dagger}, \alpha^{l-1})$ so that $\mathsf{E}(\alpha^{l-1}, \alpha^{\dagger}) = \frac{\breve{g}(\alpha^{\dagger}|y)}{\breve{g}(\alpha^{l-1}|y)}$. Here if $\mathsf{E}(\alpha^{l-1}, \alpha^{\dagger}) > 1$ the chain moves to $\alpha^{\dagger}$ unconditionally, otherwise it moves with probability given by $\frac{\breve{g}(\alpha^{\dagger}|y)}{\breve{g}(\alpha^{l-1}|y)}$. That is, we always accept the draw if we move uphill in the distribution since we want to visit areas where the density is higher. If the draw makes us move downhill, we stay at the same point with probability equal to $1 - \mathsf{E}(\alpha^{l-1}, \alpha^{\dagger})$ and explore new areas with probability equal to $\mathsf{E}(\alpha^{l-1}, \alpha^{\dagger})$. Note that, if we are already in an area with high probability, $\mathsf{E}(\alpha^{l-1}, \alpha^{\dagger})$ will be small.

As with the Gibbs sampler, a simple sufficient condition which insures that the MH algorithm converges is available and only requires some restrictions on (9.44). In fact, if for every $\alpha_0 \in A$ and every $A_1 \in A$ with $pr(\alpha \in A_1|y) > 0$, it is the case that $P(\alpha^l \in A_1|\alpha^{l-1}, y) > 0$, where P is the transition induced by (9.44), then the MH transition kernel is ergodic and its unique invariant distribution is $g(\alpha|y)$.

In implementing the MH algorithm it is therefore important to appropriately choose the transition density. One possibility is to set $\mathsf{P}(\alpha^\dagger, \alpha^{l-1}) = \mathsf{P}(\alpha^\dagger - \alpha^{l-1})$, so that the candidate draw is taken from a multivariate distribution centered at α^{l-1}. This is what we call the random walk version of the MH algorithm : $\alpha^\dagger = \alpha^{l-1} + v$. To get "reasonable" acceptance rates we need to adjust carefully Σ_v and the choice is application dependent.

Alternatively, one could use the independent chain version of the algorithm where $\mathsf{P}(\alpha^\dagger, \alpha^{l-1}) = \mathsf{P}(\alpha^\dagger)$, in which case $\mathsf{E}(\alpha^{l-1}, \alpha^\dagger) = min[\frac{\breve{g}(\alpha^\dagger|y)\mathsf{P}(\alpha^\dagger)}{\breve{g}(\alpha^{l-1}|y)\mathsf{P}(\alpha^{l-1})}, 1]$. If this alternative is chosen, both the location and the shape of $\mathsf{P}(\alpha^\dagger)$ need to be monitored to insure reasonable acceptance rates.

The independent chain version of the MH algorithm shares features with both acceptance and importance sampling. However, while the latter two approaches place a low probability of acceptance (or a low weight) on a draw that is far away from the posterior, the independent chain assigns a low probability of accepting the candidate draw if the weighted ratio of the kernels at the previous and current draw is low, where the weight is $\frac{\mathsf{P}(\alpha^\dagger)}{\mathsf{P}(\alpha^{l-1})}$.

In general, one needs to make sure that the algorithm avoids excessively high or excessively low acceptance rates since, in the first case, the exploration of the posterior is slow and, in the second, a large region with high posterior probability is left undersampled. An acceptance rate of 35-40% should be considered good.

Example 9.36 *In example 9.29 we drew* $\alpha^\dagger \sim \mathsf{N}(\tilde{\alpha}, \tilde{\Sigma}_\alpha)$ *and accepted the draw if* $\alpha^\dagger > 0$*, so that the probability of acceptance is* $(2\pi)^{-0.5}|\Sigma_\alpha|^{0.5} \int_{\alpha>0} exp[-0.5(\alpha - \tilde{\alpha})'\tilde{\Sigma}_\alpha^{-1}(\alpha - \tilde{\alpha})]d\alpha$*. If this probability is too small, the algorithm is impractical. Suppose instead we draw from* $\mathsf{P}(\alpha^\dagger, \alpha^{l-1}) = (2\pi)^{-0.5k}|\Sigma^\dagger|^{-0.5}exp[-0.5(\alpha^\dagger - \alpha^{l-1})'(\Sigma^\dagger)^{-1}(\alpha^\dagger - \alpha^{l-1})]$*, where* $\Sigma^\dagger$ *is the variance of the shocks. If* $\Sigma^\dagger$ *too small, a large number of draws is needed to cover the set where* $\alpha > 0$*. If it is too large, a large number of draws for* $\alpha^\dagger$ *will be negative. Hence, to insure an appropriate coverage of the posterior,* $\Sigma^\dagger$ *has to be carefully selected.*

Example 9.37 *Consider a bivariate normal distribution for* $z = (x, y)$ *with mean* $(1, 2)$*, variances equal to 1 and covariance equal to 0.8. A scatter plot (using 4000 draws) obtained by simulating (x,y) from this distribution is in the first box of the left hand side column of figure 9.6: it is easy to see that the ellipsoids are very thin and positively inclined. To approximate this distribution we use a MH algorithm with a reflecting random walk transition* $(z^\dagger - \bar{z}) = (z^{l-1} - \bar{z}) + v$ *where the incremental variable* v *is uniformly distributed in the interval* $[-0.5, 0.5]$ *for both coordinates. Here, the probability of accepting the draw is equal to* $\min(\frac{\exp[-0.5(z^\dagger - \bar{z})\Sigma^{-1}(z^\dagger - \bar{z})]}{\exp[-0.5(z^{l-1} - \bar{z})\Sigma^{-1}(z^{l-1} - \bar{z})]}, 1)$*. We also consider a Gibbs sampler, which uses the conditional distributions* $(x|y)$ *and* $(y|x)$*, given by* $(x|y) \sim \mathsf{N}(1 + \rho(y - 2), 1 - \rho^2)$*;* $(y|x) \sim$

$\mathsf{N}(2+\rho(x-1), 1-\rho^2)$ *where* ρ *is the correlation coefficient, to produce a sample from the posterior. The second and third rows of figure 9.6 present a sample of 4000 draws from the posteriors obtained with these two simulators. Both approaches approximate reasonably well the target. The Gibbs sampler is slightly superior but the acceptance rate of the MH algorithm is high (55%) and the tail of the distribution are not fully explored. Better acceptance rates would probably lead to a better covering of the target distribution. Note also the sticking similarities in the estimates of the marginal of* x *(see second column of figure 9.6).*

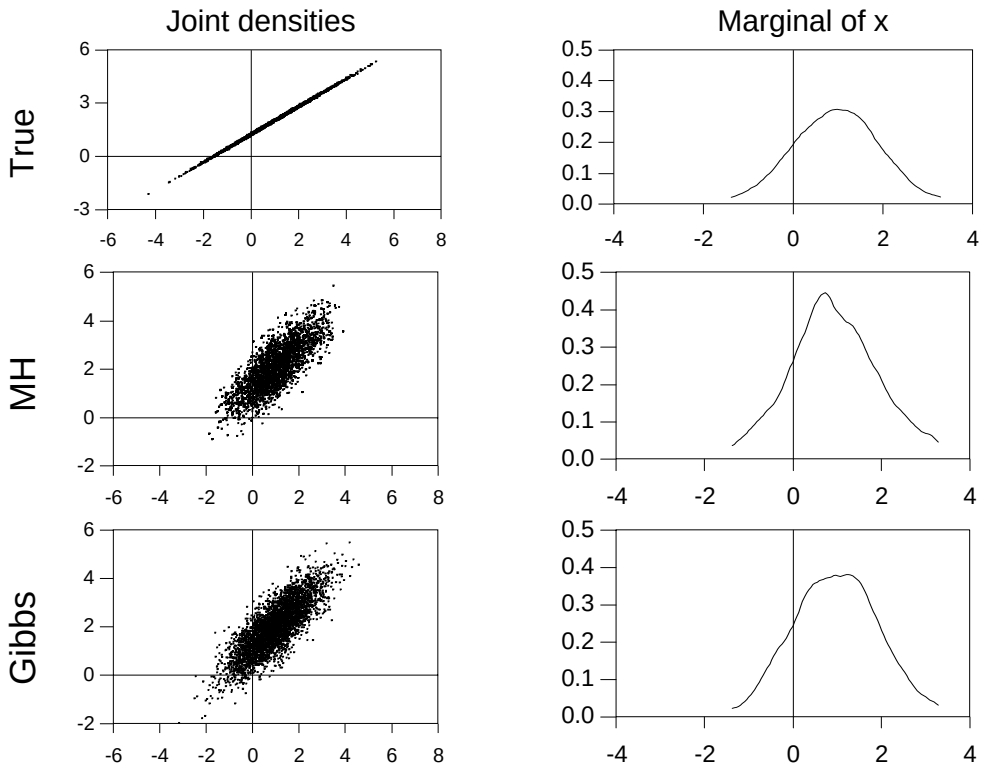


Figure 9.6: MCMC simulations

Before we close this section we would like to emphasize few important features of MCMC methods. First, Markov Chain methods work better than importance or acceptance sampling because the distribution from which draws are taken changes at each iteration (see figure 9.3). Therefore, the transition kernel is time dependent. Second, although these methods are based on Markov chains, the Markov property can be dispensed of. For example, the chain may depend on the whole history of draws. Still, if the sufficiency conditions for convergence are satisfied, both the Gibbs and the MH samplers will generate a sequence

from the posterior. Third, it is important to stress that Markov chain methods produce correlated draws. Therefore, either we eliminate this dependence via skip-sampling or posterior standard errors need to be adjusted to take this feature into account.

Exercise 9.30 *Using the setup of example 9.37, vary cov(x, y) from 0.1 to 0.9 in increments of 0.2. Show the values of Σ_B and Σ_W, when you start from $\alpha_1 = \pm 2.5; \alpha_2 = \pm 2.5$, and the interquantile range of the simulated distribution for samples of 100, 500, 1000 elements. Show the scatter plots obtained with the Gibbs sampler and the MH algorithm.*

Exercise 9.31 *Consider the model $y_t = \alpha y_{t-1} + e_t$, $e_t \sim \mathsf{N}(0, \sigma_e^2)$. The density of $(y_1, \ldots, y_t)$ is $f(y|\alpha, \sigma_e^2) \propto (\sigma_e^2)^{0.5(T-1)} \exp\{-0.5\sigma_e^{-2} \sum_{t=2}^{T}(y_t - \alpha y_{t-1})^2\} \times \exp\{-0.5\sigma_e^{-2} y_1^2(1-\alpha^2)\} [\sigma_e^{-2}(1-\alpha^2)]^{-0.5}$, where the last term is the density of y_1. Suppose the only available prior information is $\alpha < 1$. Show the form of the posterior for (α, σ_e^2) and describe how to use a MH algorithm to sample from it.*

9.6 Robustness

Whenever the prior distribution has subjective features and/or the sample is small, it is important to know how sensitive are posterior outcomes to the choice of prior distributions. Robustness is crucial, for example, in evaluating the quality of a DSGE model. Since prior distributions are conveniently chosen so as to make calculations simple, and typically centered around calibrated values, it is imperative to verify that posterior inference does not depend on the form of the prior distributions nor on its spread.

A way to assess the robustness of the posterior conclusions is to select an alternative prior density $g_1(\alpha)$, with support included in $g(\alpha)$, and use it to reweight posterior draws. Let $w(\alpha) = \frac{g_1(\alpha)}{g(\alpha)}$. Then $E_1[h(\alpha)] = \int h(\alpha) g_1(\alpha) d\alpha = \int h(\alpha) w(\alpha) g(\alpha) d\alpha$ so that $h_1(\alpha) \approx \frac{\sum_l w(\alpha^l) h(\alpha^l)}{\sum_l w(\alpha^l)}$.

Example 9.38 *Continuing with exercise 9.27, suppose $g(\alpha)$ is normal with mean 1 and variance 10. Then $g(\alpha|y)$ is normal with mean $\tilde{\alpha} = \tilde{\Sigma}_\alpha(0.1 + \sigma_e^{-2} x' x \alpha_{ols})$ and variance $\tilde{\Sigma}_\alpha = (0.1 + \sigma_e^{-2} x'x)^{-1}$. If one wishes to examine how forecasts produced by the model change if the prior variance is reduced (for example, to 5) two alternatives are possible: (i) draw a sequence α_1^l from a normal posterior with mean $\Sigma_\alpha^{-1}(0.2 + \sigma^{-2} x' x \alpha_{ols})$ and variance $\Sigma_\alpha^{-1} = (0.2 + \sigma^{-2} x'x)^{-1}$, compute forecasts for each α_1^l and compare results; (ii) weight each original posterior draw with $w(\alpha)$, i.e. calculate $\alpha_1^l = \frac{g_1(\alpha^l)}{g(\alpha^l)} \alpha^l$; compute forecast and compare results, $l = 1, 2, \ldots, L$.*

Exercise 9.32 *Suppose you are interested in comparing the welfare costs of a certain policy in a model like the one considered in example 9.11 when you use a non-informative prior for α and a sequence of informative priors, which are characterized by smaller and smaller prior variances. How would you check robustness here? What ingredients do you need to report to allow the reader to reweigh your results according to his/her own prior preferences?*

9.7 Estimating Returns to scale: Spain (1979-1999)

In this section we use Bayesian methods to measure the magnitude of the returns to scale parameter in the aggregate Spanish production function. The literature dealing with production function estimation is vast and cannot be summarized here. The exercise we conduct lies in the growth-accounting branch where output changes are explained by input changes and all unexplained variations are lumped together under the total factor productivity (TFP) label. We assume that output is produced with capital and labor using a Cobb-Douglass specification and split the residual into four components: a constant, a time trend, a term capturing efficiency improvements and one measurement errors. Hence $Y_t = (\alpha_0 t^{\alpha_1} \zeta_t e_t) K_t^{\alpha_2} N_t^{\alpha_3}$, where K and N are measured capital and labor inputs, ζ_t captures efficiency improvements, e_t is a multiplicative measurement error, α_0 a constant and t a time trend. Taking natural logarithms and linearly detrending leads to:

$$y_t = \alpha_2 \ln K_t + \alpha_3 \ln N_t + \ln \zeta_t + \ln e_t \tag{9.45}$$

where $y_t = \ln Y_t - \alpha_1 \ln t - \ln \alpha_0$. If we set $v_t = \ln \zeta_t + \ln e_t$, (9.45) is a standard regression model with two explanatory variables. The composite error term v_t is likely to be serially correlated but, as a first step, we neglect this possibility and assume $v_t \sim \mathsf{N}(0, \sigma_v^2)$. Let $x_t = (\ln K_t, \ln N_t)$ and $\alpha = (\alpha_2, \alpha_3)$.

Consider first a normal approximation. Using the results of example 9.9 we know that the marginal posterior for α has a multivariate t-format. We will take a normal approximation to each of the two components separately, centered at the mode α^* with curvature equal to $\Sigma(\alpha^*)$. The first row of table 9.1 gives percentiles of the posterior of the returns to scale parameter obtained using this approximation: the interquartile range is small and the median of the posterior is only 0.39, suggesting the presence of strong decreasing returns to scale. Posterior estimates of $(\alpha_2, \alpha_3, \sigma_v^2)$ can also be obtained with the Gibbs sampler. Suppose that $\alpha =\sim \mathsf{N}(\bar{\alpha}, \bar{\Sigma}_\alpha)$ and that $\sigma_v^{-2} \sim \mathsf{G}(a_1, a_2)$ where $\bar{\alpha} = (0.4, 0.6)$, $\bar{\Sigma}_\alpha = \text{diag}(0.05, 0.05)$ and we let $a_1 = a_2 = 10^{-5}$, to make the prior on σ_v^{-2} non-informative. The two conditional distributions are $(\alpha|\sigma_v^2, y, x) \sim \mathsf{N}(\tilde{\alpha}, \tilde{\Sigma})$ where $\tilde{\Sigma}_\alpha = (\bar{\Sigma}_\alpha^{-1} + \sigma_v^{-2} x'x)^{-1}$, $\tilde{\alpha} = \tilde{\Sigma}_\alpha(\bar{\Sigma}_\alpha^{-1}\bar{\alpha} + \sigma_v^{-2} x' x a_{OLS})$, where a_{OLS} is the OLS estimator of α and $(\sigma_v^{-2}|\alpha, y, x) \sim \mathsf{G}(a_1 + 0.5 * T, a_2 + 0.5 * (y - x\alpha_{ols})'(y - x\alpha_{ols}))$. Posterior distributions are obtained discarding the initial 500 draws and keeping one every 50 of the next 5000 draws. The second row of the table shows that the interquartile range of the returns to scale parameter is smaller than the one obtained with a normal distribution and the median is slightly higher. However, also with this approximation, decreasing returns are strong.

Next, we allow serial correlation in v_t. We assume $\rho(\ell) v_t = \epsilon_t$, so that model (9.45) is transformed into $y_t^0 \equiv \rho(\ell) y_t = \rho(\ell) x_t \alpha + \epsilon_t = x_t^0 \alpha + \epsilon_t$. The situation is now identical to the previous one, except that a new vector of parameters $\rho(\ell)$ needs to be estimated. Let $\rho(\ell) = 1 - \rho\ell$ and take the first observation to be fixed. The likelihood function is $\mathcal{L}(\alpha, \rho|y, x) = (\sigma_e^2)^{-0.5(T-1)} \exp\{-\frac{1}{\sigma_e^2} \sum_{t=2}^{T} (y_t^0 - x_t^0\alpha)^2\}$. Assume the same prior for α; let the prior for σ_ϵ^{-2} to be $\mathsf{G}(0.5, 0.5)$ and let $g(\rho)$ be normal, centered at 0.8, with variance

0.1, truncated outside the range (-1, 1) i.e. $\rho \sim \mathsf{N}(0.8, 0.1) \times \mathcal{I}_{[-1,1]}$. The conditional posterior for α is identical to the one previously derived and the one for σ_ϵ^{-2} has the same format as the one for σ_v^{-2}. The conditional distribution for ρ is normal with mean $\tilde{\rho} = \tilde{\Sigma}_\rho(8+\sigma_\epsilon^{-2}V'v)$ and variance $\tilde{\Sigma}_\rho = (10+\sigma^{-2}V'V)^{-1}$, truncated outside the range (-1,1) where $v = (v_2, \ldots, v_T)' = y - \alpha_2 \ln K - \alpha_3 \ln N$, and V is a $(T-1) \times 1$ vector with $j-th$ element given by v_{t-j-1}. Drawing from this distribution is easy since we have taken the first observation as given. Had we not done that, the computation of the conditional posterior for ρ would have required a pass with a MH algorithm. Serial correlation is important: the median value of the posterior of ρ is 0.86 and the interquartile range is (0.84, 0.90). However, returns to scale estimates are similar to the previous ones; only the estimate of the upper 75-th percentile is slightly larger (see third row of table 9.1).

	Percentiles		
Method	25th	50th	75th
Normal approximation	0.35	0.39	0.47
Basic Gibbs	0.36	0.41	0.44
Gibbs with AR(1) errors	0.35	0.41	0.48
Gibbs with latent variable	0.33	0.41	0.45

Table 9.1: Posterior distribution of returns to scale

It is clear that ζ_t, apart from technological progress, includes everything which is excluded from the production function; for example, public capital or human capital. One way of thinking about these influences is to treat ζ_t as a latent variable and let $\zeta_t = \delta z_t + \epsilon_t^\zeta$ where z_t are observables (in our case z_t measures public capital) and ϵ_t^ζ represents the true technological progress. With this specification, the model has a hierarchical latent variable structure: conditional on (x_t, ζ_t), y_t is normal with mean $x_t\alpha + \delta z_t$ and variance $\sigma_e^2 + \sigma_\zeta^2$; conditional on (y_t, x_t), ζ_t is normal with mean δz_t and variance σ_ζ^2. The specification has two new parameters σ_ζ^2 and δ. We let $\delta \sim \mathsf{N}(\bar{\delta}, \bar{\Sigma}_\delta)$ and set $\bar{\delta} = 0$ and $\bar{\Sigma}_\delta = 0.5$. Since σ_ζ^2 and σ_e^2 can not be identified separately with the short available data, we set $\sigma_\zeta^2 = \sigma_e^2$. The conditional posterior for δ is normal with mean $\tilde{\delta} = \tilde{\Sigma}_\delta(\bar{\Sigma}_\delta^{-1}\bar{\delta} + 0.5\sigma_e^{-2}z'\zeta_t)$ and variance $\tilde{\Sigma}_\delta = (\bar{\Sigma}_\delta^{-1} + 0.5\sigma_e^{-2}z'z)^{-1}$. Since the posterior distribution for δ is centered around zero (median -0.0004, interquartile range (-0.003, 0.004)), there is little evidence that ζ_t is influenced by public capital. Hence, the shape of the posterior distribution of the returns to scale parameter is roughly unchanged (see fourth row of table 9.1).

There are many extensions one could consider to refine these estimates. For example, we could think that measured inputs are different from effective ones and let e.g. $N_t = \exp\{a_N z_{Nt}\}N_{1t}$, where z_{Nt} are factors which affect the efficiency of measured input, such as education, unionization, etc. In cross sectional comparisons, this refinement could be important. N_t now becomes a latent variable and a_N an additional set of parameters which can be estimated once z_{Nt} are specified. Note that, in this case, the model has a bilinear form, but the posterior distribution of the parameters can still be obtained with the Gibbs

sampler, as shown by Koop, Osiewalski, Steel (2000).

Chapter 10: Bayesian VARs

We have seen in chapter 4 that VAR models can be used to characterize any vector of time series under a minimal set of conditions. We have also seen that since VARs are reduced form models, identification restrictions, motivated by economic theory, are needed to conduct meaningful policy analysis. Reduced form VARs are also typically unsuitable for forecasting out-of-sample. To reasonably approximate the Wold representation it is in fact necessary to have a VAR with long lags. A generous parametrization means that unrestricted VARs are not operational alternatives to either standard macroeconometric models, where insignificant coefficients are purged out of the specification, or to parsimonious time series models since, with a limited number of degrees of freedom, estimates of VAR coefficients are imprecise and forecasts have large standard errors.

It is useful to think of the construction of an empirical model as the process of combining historical and a-priori information, both of statistical and of economic nature. Alternative modeling techniques provide different a-priori information or different relative weights to sample and prior information. Unrestricted VARs employ a-priori information very sparsely - in choosing the variables of the VAR; in selecting the lag length of the model; in imposing identification restrictions. Because of this choice, overfitting may obtain when the data set is short, sample information is weak or the number of parameters is large. In-sample overfitting typically translates into poor forecasting performance, both in unconditional and conditional sense. Bayesian methods can solve these problems: they can make in-sample fitting less dramatic and improve out-of-sample performance. While Bayesian VAR (BVAR) were originally devised to improve macroeconomic forecasts, they have evolved dramatically and they are used now for a variety of purposes.

This chapter describes Bayesian methods for a variety of VAR models. First, we present the decomposition of the likelihood function of a VAR and the construction of the posterior distribution for a number of prior specifications. We also show the link between posterior mean estimates and classical estimates obtained when the coefficients of the VAR model are subject to stochastic linear constraints. The third section describes Bayesian structural VARs and block recursive structures which arise e.g., in models with some exogenous variables or in two country models with (overidentifying) linear restrictions on the contemporaneous impact of the shocks. The fourth section, discusses time varying BVAR models. These models have a state space representation and this helps in constructing both empirical Bayes and fully hierarchical posterior estimates of the VAR coefficients and of

the covariance matrix. We show that these structures generate a variety of distributional patterns and can be used to model series with thick tails, with smoothly evolving pattern, or displaying coefficients switching over a finite number of states.

The fifth section deals with multiple BVAR models: these structures are becoming popular in empirical practice, for example, when comparing the effects of monetary policy shocks in different countries or the growth behavior in different regions, and present interesting complications relative to single unit BVAR models. We show how to obtain posterior estimates of the coefficients of the model for each unit and how to obtain estimates of the mean effect across units, which often is the center of interest for applied investigators. We also describe a procedure to endogenously group units with similar characteristics. This is useful when one wants to distinguish the impact of certain shocks on e.g. small or large firms, or when policy advice requires some particular endogenous classifications (e.g. income per-capita, education level, indebtness, etc.). The last part of the section studies Bayesian Panel VAR models with cross unit interdependencies. These models are suited to study e.g., the transmission of shocks across countries or the effects of increased interdependencies in various world economies. Because of the large number of parameters, it is impossible to estimate them with classical methods and suitable (prior) restrictions need to be imposed for estimation. With such a respecification, these models are easily estimable with Monte Carlo Markov Chain methods.

Since the chapter deals with models of increasing complexity, increasingly complex methods will be used to compute posteriors. The techniques described in chapter 9 are handy here: conjugate priors allow the derivation of analytic forms for the conditional posteriors; Markov Chain Monte Carlo methods are used to draw sequences from the posterior distributions.

10.1 The Likelihood function of an m variable VAR(q)

Throughout this chapter we assume that the VAR has the form $y_t = A(L)y_{t-1} + C\bar{y}_t + e_t,\ e_t \sim (0, \Sigma_e)$, where y_t includes m variables, each of which has q lags, while the constant and other deterministic variables (trends, seasonal dummies) are collected into the $m_c \times 1$ vector $\bar{y}_t$. Hence, the number of regressors in each equation is $k = mq + m_c$ and there are mk coefficients in the VAR.

Following the steps described in chapter 4, we can rewrite the VAR in two alternative formats, both of which will be used in this chapter:

$$\mathsf{Y} = \mathsf{X}\mathsf{A} + \mathsf{E} \tag{10.1}$$

$$y = (I_m \otimes \mathsf{X})\alpha + e \quad e \sim (0, \Sigma_e \otimes I_T) \tag{10.2}$$

where Y and E are $T \times m$ matrices and X is a $T \times k$ matrix, $\mathsf{X}_t = [y'_{t-1}, \ldots, y'_{t-q}, \bar{y}'_t]$; y and e are $mT \times 1$ vectors, I_m is the identify matrix of dimension m, and $\alpha = vec(\mathsf{A})$ is a $mk \times 1$ vector. Using (10.2) the likelihood function is

$$\mathcal{L}(\alpha, \Sigma_e) \propto |\Sigma_e \otimes I_T|^{-0.5} \exp\{-0.5(y - (I_m \otimes \mathsf{X})\alpha)'(\Sigma_e^{-1} \otimes I_T)(y - (I_m \otimes \mathsf{X})\alpha)\} \tag{10.3}$$

To derive a useful decomposition of (10.3) note that

$$\begin{aligned}(y-(I_m\otimes \mathsf{X})\alpha)'(\Sigma_e^{-1}\otimes I_T)(y-(I_m\otimes \mathsf{X})\alpha) &= \\ (\Sigma_e^{-0.5}\otimes I_T)(y-(I_m\otimes \mathsf{X})\alpha)'(\Sigma_e^{-0.5}\otimes I_T)(y-(I_m\otimes \mathsf{X})\alpha) &= \\ [(\Sigma_e^{-0.5}\otimes I_T)y-(\Sigma_e^{-0.5}\otimes \mathsf{X})\alpha)]'[(\Sigma_e^{-0.5}\otimes I_T)y-(\Sigma_e^{-0.5}\otimes \mathsf{X})\alpha)]\end{aligned}$$

Also $(\Sigma_e^{-0.5}\otimes I_T)y-(\Sigma_e^{-0.5}\otimes \mathsf{X})\alpha = (\Sigma_e^{-0.5}\otimes I_T)y-(\Sigma_e^{-0.5}\otimes \mathsf{X})\alpha_{ols}+(\Sigma_e^{-0.5}\otimes \mathsf{X})(\alpha_{ols}-\alpha)$ where $\alpha_{ols}=(\Sigma_e^{-1}\otimes \mathsf{X}'\mathsf{X})^{-1}(\Sigma_e^{-1}\otimes \mathsf{X})'y$. Therefore:

$$\begin{aligned}(y-(I_m\otimes \mathsf{X})\alpha)'(\Sigma_e^{-1}\otimes I_T)(y-(I_m\otimes \mathsf{X})\alpha) &= \\ ((\Sigma_e^{-0.5}\otimes I_T)y-(\Sigma_e^{-0.5}\otimes \mathsf{X})\alpha_{ols})'((\Sigma_e^{-0.5}\otimes I_T)y-(\Sigma_e^{-0.5}\otimes \mathsf{X})\alpha_{ols}) &+ \quad (10.4)\\ (\alpha_{ols}-\alpha)'(\Sigma_e^{-1}\otimes \mathsf{X}'\mathsf{X})(\alpha_{ols}-\alpha) & \quad (10.5)\end{aligned}$$

The term in (10.4) is independent of α and looks like a sum of squared errors. The one in (10.5) looks like the scaled square error of α_{ols}. Putting the pieces back together we have:

$$\begin{aligned}\mathcal{L}(\alpha,\Sigma_e) &\propto |\Sigma_e\otimes I_T|^{-0.5}\exp\{-0.5(\alpha-\alpha_{ols})'(\Sigma_e^{-1}\otimes \mathsf{X}'\mathsf{X})(\alpha-\alpha_{ols}) \\ &- 0.5[(\Sigma_e^{-0.5}\otimes I_T)y-(\Sigma_e^{-0.5}\otimes \mathsf{X})\alpha_{ols})'[(\Sigma_e^{-0.5}\otimes I_T)y-(\Sigma_e^{-0.5}\otimes \mathsf{X})\alpha_{ols})]\} \\ &= |\Sigma_e|^{-0.5k}\exp\{-0.5(\alpha-\alpha_{ols})'(\Sigma_e^{-1}\otimes \mathsf{X}'\mathsf{X})(\alpha-\alpha_{ols})\} \\ &\times |\Sigma_e|^{-0.5(T-k)}\exp\{-0.5tr[(\Sigma_e^{-0.5}\otimes I_T)y \\ &- (\Sigma_e^{-0.5}\otimes \mathsf{X})\alpha_{ols})'(\Sigma_e^{-0.5}\otimes I_T)y-(\Sigma_e^{-0.5}\otimes \mathsf{X})\alpha_{ols})]\} \\ &\propto \mathbb{N}(\alpha|\alpha_{ols},\Sigma_e,\mathsf{X},y)\times\mathbb{W}(\Sigma_e^{-1}|y,\mathsf{X},\alpha_{ols},T-k-m-1) \quad (10.6)\end{aligned}$$

where tr is the trace of a matrix. The likelihood function of a VAR(q) can therefore be decomposed into the product of a Normal density for α, conditional on the OLS estimate α_{ols} and on Σ_e, and a Wishart density for Σ_e^{-1}, conditional on α_{ols}, with scale matrix $[(y-(I_m\otimes \mathsf{X})\alpha_{ols})'(y-(I_m\otimes \mathsf{X})\alpha_{ols})]^{-1}$, and $(T-k-m-1)$ degrees of freedom (see the Appendix for the form of various distributions).

Hence, under appropriate conjugate prior restrictions, we can analytically derive the conditional posterior distribution for the VAR coefficients and the covariance matrix of the reduced form shocks. As we have seen in chapter 9, a Normal-Wishart prior conjugates the two blocks of the likelihood. Therefore, under these assumptions, the conditional posterior for α will be Normal and the conditional posterior of Σ_e^{-1} will be Wishart. Other prior assumptions on α and Σ_e also allow analytical computation of conditional posteriors. We examine them in the next section.

10.2 Priors for VARs

In this section we consider four alternative types of prior specification:

1. A Normal prior for α with Σ_e fixed.

2. A non-informative prior for both α and Σ_e.

3. A Normal prior for α, and a non-informative prior for Σ_e.

4. A Conditionally conjugate prior, i.e. a Normal for α, and a Wishart for Σ_e^{-1}.

We examine in details the derivation of the posterior distribution for the VAR coefficients for case 1. Let the prior be $\alpha = \bar{\alpha} + v_a, \quad v_a \sim N(0, \bar{\Sigma}_a)$, with $\bar{\Sigma}_a$ fixed. Then

$$\begin{aligned} g(\alpha) &\propto |\bar{\Sigma}_a|^{-0.5} exp[-0.5(\alpha-\bar{\alpha})'\bar{\Sigma}_a^{-1}(\alpha-\bar{\alpha})] \\ &\propto |\bar{\Sigma}_a|^{-0.5} exp[-0.5(\bar{\Sigma}_a^{-0.5}(\alpha-\bar{\alpha}))'\bar{\Sigma}_a^{-0.5}(\alpha-\bar{\alpha})] \end{aligned} \tag{10.7}$$

Let $\mathcal{Y} = [\bar{\Sigma}_a^{-0.5}\bar{\alpha}, (\Sigma_e^{-0.5} \otimes I_T)y]'$; $\mathcal{X} = [\bar{\Sigma}_a^{-0.5}, (\Sigma_e^{-0.5} \otimes \mathsf{X})]'$. Then:

$$\begin{aligned} g(\alpha|y) &\propto |\bar{\Sigma}_a|^{-0.5} \exp\{-0.5(\bar{\Sigma}_a^{-0.5}(\alpha-\bar{\alpha}))'\bar{\Sigma}_a^{-0.5}(\alpha-\bar{\alpha})\} \times |\Sigma_e \otimes I_T|^{-0.5} \\ &\times \exp\{(\Sigma_e^{-0.5} \otimes I_T)y - (\Sigma_e^{-0.5} \otimes \mathsf{X})\alpha)'(\Sigma_e^{-0.5} \otimes I_T)y - (\Sigma_e^{-0.5} \otimes \mathsf{X})\alpha)\} \\ &\propto \exp\{-0.5(\mathcal{Y} - \mathcal{X}\alpha)'(\mathcal{Y} - \mathcal{X}\alpha)\} \\ &\propto \exp\{-0.5(\alpha-\tilde{\alpha})'\mathcal{X}'\mathcal{X}(\alpha-\tilde{\alpha}) + (\mathcal{Y} - \mathcal{X}\tilde{\alpha})'(\mathcal{Y} - \mathcal{X}\tilde{\alpha})\} \end{aligned} \tag{10.8}$$

where

$$\tilde{\alpha} = (\mathcal{X}'\mathcal{X})^{-1}(\mathcal{X}'\mathcal{Y}) = [\bar{\Sigma}_a^{-1} + (\Sigma_e^{-1} \otimes \mathsf{X}'\mathsf{X})]^{-1}[\bar{\Sigma}_a^{-1}\bar{\alpha} + (\Sigma_e^{-1} \otimes \mathsf{X})'y] \tag{10.9}$$

Since Σ_e and $\bar{\Sigma}_a$ are fixed, the second term in (10.8) is a constant independent of α and

$$g(\alpha|y) \propto exp[-0.5(\alpha-\tilde{\alpha})'\mathcal{X}'\mathcal{X}(\alpha-\tilde{\alpha})] \propto exp[-0.5(\alpha-\tilde{\alpha})'\tilde{\Sigma}_a^{-1}(\alpha-\tilde{\alpha})] \tag{10.10}$$

Hence, the posterior density of α is Normal with mean $\tilde{\alpha}$ and variance $\tilde{\Sigma}_a = [\bar{\Sigma}_a^{-1} + (\Sigma_e^{-1} \otimes \mathsf{X}'\mathsf{X})]^{-1}$. For (10.10) to be operational we need $\bar{\Sigma}_a$ and Σ_e. Typically, $\bar{\Sigma}_a$ is arbitrarily chosen (e.g. to have a loose prior) and one uses e.g., $\Sigma_{e,ols} = \frac{1}{T-1}\sum_{t=1}^{T} e'_{t,ols}e_{t,ols}$, $e_{t,ols} = y_t - (I_m \otimes \mathsf{X})\alpha_{ols}$, in the formulas.

10.2.1 Least square under uncertain restrictions

The posterior mean for α displayed in (10.9) has the same format as a classical estimator obtained with Theil's mixed type approach when coefficients are stochastically restricted. To illustrate this point consider a univariate AR(q) with no constant:

$$\begin{aligned} \mathsf{Y} &= \mathsf{X}A + \mathsf{E} \qquad \mathsf{E} \sim (0, \Sigma_e) \\ A &= \bar{A} + v_a \qquad v_a \sim (0, \bar{\Sigma}_a) \end{aligned} \tag{10.11}$$

where $A = [A_1, \ldots A_q]'$, $\mathsf{X}_t = [y_{t-1}, \ldots y_{t-q}]$. Set $\mathcal{Y}_t = [\mathsf{Y}_t, \bar{A}']'$, $\mathcal{X}_t = [\mathsf{X}_t, I]'$, $E_t = [\mathsf{E}_t, v'_a]'$. Then $\mathcal{Y}_t = \mathcal{X}_t A + E_t$, where $E_t \sim (0, \Sigma_E)$, and Σ_E is assumed known. The (generalized) least square estimator is $A_{GLS} = (\mathcal{X}'\Sigma_E^{-1}\mathcal{X})^{-1}(\mathcal{X}'\Sigma_E^{-1}\mathcal{Y})$, which is identical to $\tilde{A}$, the mean of the posterior of A obtained with fixed Σ_e, fixed $\bar{\Sigma}_a$ and a Normal prior for A. There

is a simple but useful interpretation of this result. Prior restrictions on VAR coefficients can be treated as dummy observations which are added to the system of VAR equations. The posterior estimator will efficiently combine sample and prior information using their precisions as weights. Additional restrictions can be tagged on to the system in exactly the same fashion and posterior estimates can be obtained by combining the vector of prior restrictions with the data. We will exploit this feature later on, when we design restrictions intended to capture the existence of trends, seasonal fluctuations, etc.

Exercise 10.1 *(Hoerl and Kennard) Suppose that* $\bar{A} = 0$ *in (10.11). Show that the posterior mean of* A *is* $\tilde{A} = (\bar{\Sigma}_a^{-1} + \mathsf{X}'\Sigma_e^{-1}\mathsf{X})^{-1}(\mathsf{X}'\Sigma_e^{-1}\mathsf{Y})$. *Show that if* $\Sigma_e = \sigma_e^2 \times I_T$, $\bar{\Sigma}_a = \sigma_v^2 \times I_q$, $\tilde{A} = (I_q + \frac{\sigma_e^2}{\sigma_v^2}(\mathsf{X}'\mathsf{X})^{-1})^{-1}A_{ols}$, *where* A_{ols} *is the OLS estimator of* A.

There are two important features of exercise 10.1. First, since the restriction $\bar{A} = 0$ imposes the belief that all the coefficients are small, it is appropriate if y_t is the growth rate of financial variables like exchange rates or stock prices. Second, the last part of the exercise indicates that the posterior estimator increases the smallest eigenvalues of the data matrix by the factor $\frac{\sigma_e^2}{\sigma_v^2}$. Hence, it is useful when the $(\mathsf{X}'\mathsf{X})$ matrix is ill-conditioned (e.g. when near multi-collinearity is present).

Exercise 10.2 *Treating* $\tilde{\alpha}$ *in (10.9) as a classical estimator, show what conditions insure its consistency and its asymptotic normality.*

There is an alternative representation of the prior for case 1. Set $R\alpha = r + v_a$, $v_a \sim \mathbb{N}(0, I)$, where R is a square matrix. Then $g(\alpha)$ is $\mathbb{N}(R^{-1}r, R^{-1}R^{-1'})$ and $\tilde{\alpha} = [R'R + (\Sigma_e^{-1} \otimes \mathsf{X}'\mathsf{X})]^{-1}[R'r + (\Sigma_e^{-1} \otimes \mathsf{X})'y]$. This last expression has two advantages over (10.9). First, it does not require the inversion of the $mk \times mk$ matrix $\bar{\Sigma}_a$, which could be complicated in large scale VARs. Second, zero restrictions on some coefficients are easy to impose - in (10.9) this must be done setting some diagonal elements of $\bar{\Sigma}_a$ to infinity.

Exercise 10.3 *Using* $R\alpha = r + v_a$, $v_a \sim \mathbb{N}(0, I)$ *as a prior, show that* $\sqrt{T}(\tilde{\alpha} - \alpha_{ols}) \xrightarrow{P} 0$ *as* $T \to \infty$.

The intuition for the result of exercise 10.3 is clear: since as T grows, the importance of the data increases relative to the prior, $\tilde{\alpha}$ coincides with the unrestricted OLS estimator.

10.2.2 The Minnesota prior

The so-called Minnesota (Litterman) prior is a special case of Case 1 prior when $\bar{\alpha}$ and Σ_α are functions of a small number of hyperparameters. In particular (see, for example, RATS (2000)) this prior assumes that $\bar{\alpha} = 0$ except for $\bar{\alpha}_{i1} = 1$, $i = 1, \ldots, m$; that Σ_a is diagonal and that the $\sigma_{ij,\ell}$ element corresponding to lag ℓ of variable j in equation i has the form:

$$\sigma_{ij,\ell} = \frac{\phi_0}{h(\ell)} \qquad \text{if} \quad i = j, \forall \ell$$

$$
\begin{aligned}
&= \phi_0 \times \frac{\phi_1}{h(\ell)} \times (\frac{\sigma_j}{\sigma_i})^2 \ \text{ otherwise when } i \neq j, j \text{ endogenous}, \forall \ell \\
&= \phi_0 \times \phi_2 \ \text{ for j exogenous} \qquad (10.12)
\end{aligned}
$$

Here $\phi_i,\ i = 0, 1, 2$ are hyperparameters, $(\frac{\sigma_j}{\sigma_i})^2$ is a scaling factor and $h(\ell)$ a deterministic function of ℓ. The prior (10.12) captures features of interest to the investigator: ϕ_0 represents the tightness on the variance of the first lag; ϕ_1 the relative tightness of other variables; ϕ_2 the relative tightness of the exogenous variables and $h(\ell)$ the relative tightness of the variance of lags other than the first one. Typically, one assumes an harmonic decay $h(\ell) = \ell^{\phi_3}$(a special case of which is $h(\ell) = \ell$, a linear decay) or a geometric decay $h(\ell) = \phi_3^{-\ell+1}, \phi_3 > 0$. Since $\sigma_i, i = 1, \ldots, m$ are unknown, consistent estimates of the standard errors of the variables i, j are used in (10.12).

To understand the logic of this prior note that the m time series are a-priori represented as random walks. This specification is selected because univariate random walk models are typically good in forecasting macroeconomic time series. Note also that the random walk hypothesis is imposed a-priori: a posteriori, each time series may follow a more complicated process if there is sufficient information in the data to require it.

The variance-covariance matrix is a-priori selected to be diagonal. Hence, there is no relationship among the coefficients of various VAR equations. Moreover, the most recent lags of a variable are expected to contain more information about the variable's current value than do earlier lags. Hence, the variance of lag ℓ_2 is smaller than the variance of lag ℓ_1 if $\ell_2 > \ell_1$ for every endogenous variable of the model. Furthermore, since lags of other variables typically have less information than lags of own variables, $\phi_1 \leq 1$. Note that, if $\phi_1 = 0$, the VAR is a-priori collapsed into a vector of univariate models. Finally, ϕ_2 regulates the relative importance of the information contained in the exogenous variables and ϕ_0 controls the relative importance of sample and prior information. From (10.9) if ϕ_0 is large, prior information becomes diffuse so the posterior distribution mirrors sample information. If ϕ_0 is small, prior information dominates.

A graphical representation of this prior is in figure 10.1: all coefficients have zero prior mean (except the first own lag) and prior distributions become more concentrated for coefficients on longer lags. Moreover, the prior distributions of the lags of the variables not appearing on the left hand side of the equation are more concentrated than those of the own lags.

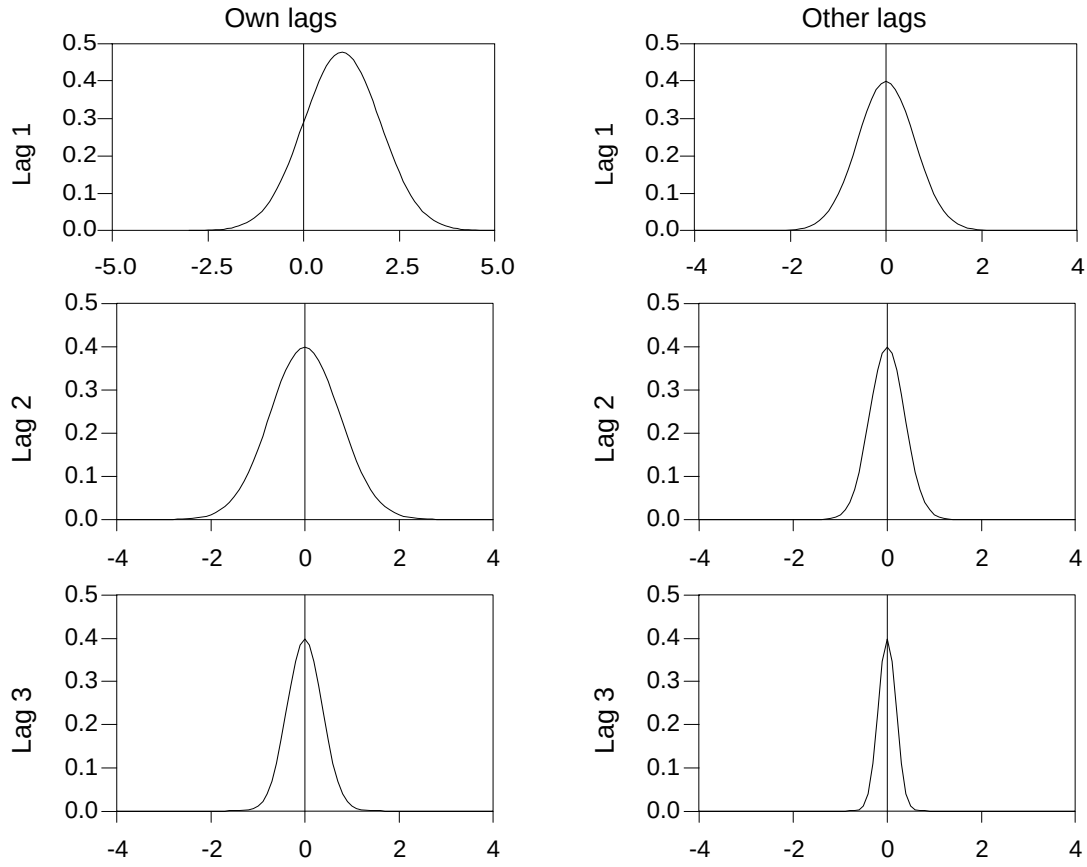


Figure 10.1: Minnesota prior.

Example 10.1 *To see what the Minnesota prior implies consider a VAR(2) with* $q = 2$, $\bar{y} = 0$, $h(\ell) = \ell$. *In this case* $\bar{\alpha} = [1, 0, 0, 0, 0, 1, 0, 0]$, *and*

$$\bar{\Sigma}_a = \begin{bmatrix} \phi_0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \phi_0\phi_1(\frac{\sigma_2}{\sigma_1})^2 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \frac{\phi_0}{2} & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \frac{\phi_0}{2}\phi_1(\frac{\sigma_2}{\sigma_1})^2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \phi_0\phi_1(\frac{\sigma_1}{\sigma_2})^2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \phi_0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \frac{\phi_0}{2}\phi_1(\frac{\sigma_1}{\sigma_2})^2 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{\phi_0}{2} \end{bmatrix}$$

There are considerable advantages in specifying $\bar{\Sigma}_a$ to be diagonal. Since the same variables appear on the right hand side of each equations, a diagonal $\bar{\Sigma}_a$ implies a diagonal $\tilde{\Sigma}_a$ so that $\tilde{\alpha}$ is the same as the vector of $\tilde{\alpha}_i$ computed equation by equation. This property is lost with other prior specifications, regardless of the assumption made on $\bar{\Sigma}_a$.

Exercise 10.4 *Using the logic of seemingly unrelated regressions show that when $g(\alpha)$ is of Minnesota type, estimating the VAR jointly gives the same posterior estimator for the coefficients of equation i as estimating each VAR equation separately.*

The dimension of α for moderate VARs is typically large: for example, if there are 5 endogenous variables, 5 lags and a constant, $k = 26$ and a $mk = 130$. With standard macro data (say, forty years of quarterly data (T=160)), maximum likelihood estimates are unlikely to have reasonable properties. The Minnesota type makes this large number of coefficients depend on a smaller vector of hyperparameters. If these are the objects estimated from the data, a better precision is expected because of the sheer dimensionality reduction (the noise to signal ratio is smaller; the number of data points per parameter increased), and out-of-sample forecasts can be improved. Note that even when the prior is false, in the sense that it does not reflect well sample information, this approach may reduce the MSE of the estimates. A number of authors have shown that VARs with a Minnesota prior produce superior forecasts to those of, say, univariate ARIMA models or traditional multivariate simultaneous equations (see e.g. Robertson and Tallman (1999) for a recent assessment). Therefore, it is not surprising that BVARs are routinely used for short-term macroeconomic forecasting in Central Banks and international institutions.

It is useful to contrast the Minnesota approach and other methods used to deal with the "curse of dimensionality". In classical approaches, "unimportant" lags are purged from the specification using t-test or similar procedures (see e.g. Favero (2001)). This approach therefore imposes strong a-priori restrictions on what variables and which lags should be in the VAR. However, dogmatic restrictions are unpalatable because they are hard to justify on both economic and statistical grounds. The Minnesota prior introduces restrictions in a flexible way: it imposes probability distributions on the coefficients of the VAR which reduce the dimensionality of the problem and, at the same time, give a reasonable account of the uncertainty faced by an investigator.

The choice of $\phi = (\phi_0,\ \phi_1,\ \phi_2,\ \phi_3)$ is important since if the prior is too loose, overfitting is hard to avoid; while if it is too tight, the data is not allowed to speak. There are three approaches one can use. In the first two, one obtains estimates of ϕ and plug-in these estimates into the expression for $\bar{\alpha}$ and $\bar{\Sigma}_a$. Then the posterior distribution of α can be obtained from (10.9) in an Empirical Bayes fashion, conditional on the ϕ estimates. In the third approach, one treats ϕ as random, assumes a prior distribution and computes fully hierarchical posterior estimates of α. To do this we need MCMC methods. For now we focus on the first two methods.

One way to choose ϕ is to use simple rules of thumb or experience. The RATS manual (2000), for example, suggests as default values $\phi_0 = 0.2,\ \phi_1 = 0.5, \phi_2 = 10^5$, an harmonic

specification for $h(\ell)$ with $\phi_3 = 1$ or 2, implying a relatively loose prior on the VAR coefficients and an uninformative prior for the exogenous variables. These values work reasonably well in forecasting a number of macroeconomic and financial variables and should be used as a benchmark or as starting points for further investigations.

The alternative is to estimate ϕ using the information contained in the data. In particular, the predictive density $f(\phi|y) = \int \mathcal{L}(\alpha|y,\phi)g(\alpha|\phi)d\alpha$, constructed on a training sample $(-\tau,\ldots,0)$, could be used. The next example shows how to do this in a simple model.

Example 10.2 *Suppose $y_t = Ax_t + e_t$, where A is a random scalar,$e_t \sim \mathbb{N}(0,\sigma_e^2)$; σ_e^2 known and let $A = \bar{A} + v_a$; $v_a \sim \mathbb{N}(0,\bar{\sigma}_a^2)$, $\bar{A}$ is fixed and $\bar{\sigma}_a^2 = h(\phi)^2$ where ϕ is a vector of hyperparameters. Then $y_t = \bar{A}x_t + \epsilon_t$ where $\epsilon_t = e_t + v_a x_t$ and the posterior kernel is:*

$$\breve{g}(\alpha,\theta|y) = \frac{1}{\sqrt{2\pi}\sigma_e h(\phi)}\exp\{-0.5\frac{(y-Ax)^2}{\sigma_e^2} - 0.5\frac{(A-\bar{A})^2}{h(\phi)^2}\} \tag{10.13}$$

where $y = [y_1,\ldots y_t]', x = [x_1,\ldots x_t]'$.Integrating (10.13) with respect to A we obtain

$$f(\phi|y) = \frac{1}{\sqrt{2\pi h(\phi)^2 tr|x'x| + \sigma_e^2}}\exp\{-0.5\frac{(y-\bar{A}x)^2}{\sigma_e^2 + h(\phi)^2 tr|x'x|}\} \tag{10.14}$$

which can be constructed and maximized, e.g., using the prediction error decomposition generated by the Kalman filter.

While in example 10.2 A is a scalar, the same logic applies when α is a vector.

Exercise 10.5 *Let $y_t = A(\ell)y_{t-1} + e_t$, $e_t \sim \mathbb{N}(0,\Sigma_e)$, Σ_e known, let $\alpha = vec(A_1,\ldots,A_q)'$ $= \bar{\alpha} + v_a$, $\bar{\alpha}$ known and $\bar{\Sigma}_a = h(\phi)^2$. Show $f(\phi|y)$ and its prediction error decomposition.*

Exercise 10.6 *Suppose that $\bar{A} = h_1(\phi)$ and $\bar{\Sigma}_a = h_2(\phi)$ in example 10.2. Derive the first order conditions for the optimal ϕ. Describe how to numerically find ML-II estimates of ϕ.*

We summarize the features of the posterior distribution of α and Σ_e obtained with the other three prior specifications in the next exercises (see Kadiyala and Karlsson (1997)).

Exercise 10.7 *Suppose that $g(\alpha,\Sigma_e^{-1}) \propto |\Sigma_e^{-1}|^{0.5(m+1)}$. Show that the joint posterior has a Normal-Wishart shape with $(\alpha|\Sigma_e,y) \sim \mathbb{N}(\alpha_{ols},(\Sigma_e^{-1}\otimes \mathsf{X}'\mathsf{X})^{-1})$; $(\Sigma_e^{-1}|y) \sim \mathbb{W}([(y-(I\otimes\mathsf{X})\alpha_{ols})'(y-(I\otimes\mathsf{X})\alpha_{ols})]^{-1}, T-k)$ and that $(\alpha|y)$ has a t-distribution with parameters $((y-(I\otimes\mathsf{X})\alpha_{ols})'(y-(I\otimes\mathsf{X})\alpha_{ols}),\alpha_{ols},T-k)$, where α_{ols} is the OLS estimator of α. Conclude that, a-posteriori, the elements of α are dependent (Hint: Stare at the variance of α).*

Exercise 10.8 *Suppose that the joint prior for (α,Σ_e^{-1}) is Normal-diffuse, i.e. $g(\alpha) \sim \mathbb{N}(\bar{\alpha},\bar{\Sigma}_a)$ where both $\bar{\alpha}$ and $\bar{\Sigma}_a$ are known and $g(\Sigma_e) \propto |\Sigma_e^{-1}|^{0.5(m+1)}$. Show that $g(\alpha|y) \propto \exp\{0.5(\alpha-\bar{\alpha})'\bar{\Sigma}_a^{-1}(\alpha-\bar{\alpha})\} \times |(y-(I\otimes\mathsf{X})\alpha_{ols})'(y-(I\otimes\mathsf{X})\alpha_{ols}) + (\alpha-\alpha_{ols})'(\mathsf{X}'\mathsf{X})(\alpha-$*

$\alpha_{ols})|^{-0.5T}$. *Conclude that* $g(\alpha|y)$ *is the product of the normal prior and the same t-distribution found in exercise 10.7. Argue that there is posterior dependence among equations, even when* $\bar{\Sigma}_a$ *is diagonal.*

Exercise 10.9 *Let* $g(\alpha|\Sigma_e) \sim \mathbb{N}(\bar{\alpha}, \Sigma_e \otimes \bar{\Omega})$ *and* $g(\Sigma_e^{-1}) \sim \mathbb{W}(\bar{\Sigma}^{-1}, \bar{\nu})$. *Show that* $g(\alpha|\Sigma_e, y) \sim \mathbb{N}(\tilde{\alpha}, \Sigma_e \otimes \tilde{\Omega})$, $g(\Sigma_e^{-1}|y) \sim \mathbb{W}(\tilde{\Sigma}^{-1}, T + \bar{\nu})$. *Give the form of* $\tilde{\alpha}, \tilde{\Omega}, \tilde{\Sigma}^{-1}$. *Show that* $(\alpha|y)$ *has a t-distribution with parameters* $(\tilde{\Omega}^{-1}, \tilde{\Sigma}, \tilde{\alpha}, T + \bar{\nu})$. *Assume that* $\bar{\Omega} = \text{diag}\{\bar{\omega}_{ii}\}$ *where* $\bar{\omega}_{ii}$ *is parametrized as in the Minnesota prior (except that* $\phi_1 = 1$*); suppose that* $\bar{\nu} = m + 2$ *and that* $\bar{\sigma}_{ii} = \text{diag}(\bar{\Sigma}) = (\bar{\nu} - m - 1)s_i^2$, *where* s_i^2 *is the estimated variance of* e_i. *Show that there is posterior dependence among the equations.*

10.2.3 Adding other prior restrictions

We can add a number of other statistical restrictions to the standard Minnesota prior without altering the form of the posterior moments. For example, an investigator may be interested in studying the dynamics at seasonal frequencies and therefore want to use the seasonal information to set up prior restrictions. The simplest way to deal with seasonality is to include a set of dummies in the VAR and treat their coefficients in the same way as the coefficients on the constant.

Example 10.3 *In quarterly data, a prior for a bivariate VAR(2) with four seasonal dummies has mean equal to* $\bar{\alpha} = [1, 0, 0, 0, 0, 0, 0, 0, 0|0, 1, 0, 0, 0, 0, 0, 0, 0]$ *and the block of* Σ_a *corresponding to the seasonal dummies has diagonal elements,* $\sigma_{dd} = \phi_0\phi_s$. *Here* ϕ_s *represents the tightness of the seasonal information (and a large* ϕ_s *implies little prior information).*

Seasonality, however, is hardly deterministic (in that case, it would be easy to eliminate it if we did not want it) and seasonal dummies only roughly account for seasonal variations. As an alternative, note that seasonal series display a peak (or a wide mass) in the spectrum at some or all seasonal frequencies. When a series has a peak at frequency ω_0 it must be the case that in the model $y_t = D(\ell)e_t$, $|D(\omega_0)|^2$ is large. A large $|D(\omega_0)|^2$ implies that $|A(\omega_0)|^2$ should be small, where $A(\ell) = D(\ell)^{-1}$, which in turns implies $\sum_{j=1}^{\infty} A_j cos(\omega_0 j) \approx -1$.

Example 10.4 *In quarterly data,* $\omega_0 = \frac{\pi}{2}, \pi$ *(cycles corresponding to 4 and 2 quarters) and a peak at, say,* $\frac{\pi}{2}$ *implies that* $-A_2 + A_4 - A_6 + A_8 + \ldots$ *must be close to* -1.

The same idea applies to multivariate models. Omitting constants, the MA representation is $y_t = D(\ell)e_t$ and the spectral density of y_t is $\mathcal{S}_y(\omega) = |D(\omega)|^2 \frac{\Sigma_e}{2\pi}$. Since $D(\omega) = \sum_j D_j(cos(\omega j) + isin(\omega j))$, a peak in $\mathcal{S}_y$ at ω_0 implies that $\sum_j D_j cos(\omega_0 j)$ is large and $\sum_{j=1}^{\infty} A_j cos(\omega_0 j) \approx -1$.

We can cast these restrictions in the form $R\alpha = r + v_a$, where $r = [-1, \ldots, -1]'$, R is a $m_1 \times mk$ matrix and m_1 is the number of seasonal frequencies. In quarterly data, if the first variable of the VAR displays seasonality at both $\frac{\pi}{2}, \pi$ then:

$$R = \begin{bmatrix} 0 & -1 & 0 & 1 & 0 & -1 & \ldots & 0 \\ -1 & 1 & -1 & 1 & -1 & 1 & \ldots & 0 \end{bmatrix}$$

These restrictions can be added to those of the original (Minnesota) prior and combined with the data using the logic of Theil's mixed type estimation, once Σ_{v_a} is selected. The same approach can also be used to account for the presence of peaks in other parts of the spectrum, as it is shown in the next exercise.

Exercise 10.10 *(Canova)*
(i) Show that a peak in the spectral density at frequency zero in variable i implies $\sum_{j=1}^{\infty} A_{ji} \approx -1$. *Cast this constraint in the form of an uncertain linear restriction.*
(ii) Show that a large mass in the band $(\frac{2\pi}{j} \pm \varepsilon)$, *some* j, ε *small, in variable i implies* $\sum_{j=1}^{\infty} A_{ji} cos(j\omega_0) \approx -1$, *for all* ω_0 *in the band. Cast these constraints in the form of uncertain linear restrictions.*
(iii) Show that a high coherence at $\omega_0 = \frac{\pi}{2}$ *in series i and* i' *of a VAR implies that* $\sum_{j=1}^{\infty}(-1)^j A_{i'i'(2j)} + \sum_{j=1}^{\infty}(-1)^j A_{ii(2j)} \approx -2$. *Cast this constraint in the form of an uncertain linear restriction.*

Other types of probabilistic constraints can be imposed in a similar way. As long as r, R and $\text{var}(v_a)$ are fixed, combining prior and sample information presents no conceptual difficulty: the dimensionality of R and of r changes, but the form of the posterior moments of α is unchanged.

10.2.4 Some Applied tips

There are few practical issues a researcher faces in setting-up a Minnesota prior for a VAR. First, in simple applications it is typical to use default values for the hyperparameters ϕ. While this is a good starting point, it is not clear that this choice is appropriate in all forecasting situations or when structural inference is required. In these cases, sensitivity analysis may give information about interesting local derivatives, e.g. how much the MSE of the forecasts change when ϕ varies within a small range of the default value. If differences are large, should hyperparameters be chosen to get the best out-of-sample performance? Since hyperparameters describe features of the prior they should be chosen using the predictive density. Using ex-post MSE statistics poses few operational problems. Which forecasting horizon should be chosen to select the hyperparameters? If different horizons require different parameters, how should one proceed? The use of the predictive density provides a natural answer to these questions. Since predictive densities can be decomposed into the product of one-step ahead prediction errors, hyperparameters chosen optimizing the predictive density minimize the one-step ahead prediction error in the training sample.

Second, in certain applications the defaults values of the Minnesota prior are clearly inappropriate: for example, a mean of one on the first lag for growth rates is unlikely to be useful. In others, one may want to have additional parameters controlling, e.g., the relative importance of certain variables in one equation or across equations. For example, one would expect lags of other variables to be less important when the left hand side of an equation there is a financial variable, but very important when there is a macroeconomic variable.

Alterations of the Minnesota prior in this direction do not change the form of the posterior so long as $\bar{\Sigma}_a$ is diagonal and Σ_e fixed.

Although the emphasis of this section has been on type 1 priors, all the arguments made remain valid when a general Normal-Wishart prior are used. Conditional on Σ_e the posterior for α is still normal. However, equation-by-equation computations are no longer efficient since the posterior covariance matrix obtained using the whole system is different from the covariance matrix obtained using each equation separately. For VARs with 5 or 6 variables and 4 or 5 lags, system wide calculations are not computationally demanding, given existing computer technology. For larger scale models such as the one of Leeper, Sims and Zha (1996), intelligent choices for the prior may dramatically simplify the computations.

How do one selects the variables to be included in a BVAR? Using the same logic described in chapter 9, specifications with different variables can be treated as different models. Therefore, a posterior odds ratio or the Leamer's version of it can be used to select the specification that best fit the data in a training sample. Consequently, one chooses the specification with the smallest one-step ahead prediction error will be preferred. Such calculations can be performed both in nested and non-nested models.

Example 10.5 *(Forecasting inflation) We use a BVAR with a Minnesota prior to forecast inflation rates in Italy. The features of inflation rates have changed dramatically in the 90's all over the world and in Italy in particular. In fact, while the autocovariance function displays remarkable persistence in the 80's (AR(1) coefficient equals 0.85), it decays pretty quickly in the 90's (AR(1) coefficient equals 0.48). In this situation, using 1980's data to choose a model or its hyperparameters may severely impair its ability to forecast in the 90's. As a benchmark for comparison we use a univariate ARIMA model, chosen using standard Box-Jenkins methods, and a three variable unrestricted VAR, including the annualized three month inflation, the unemployment rate and the annualized three month rent inflation, each with four lags. These variables were chosen among a set of ten candidates using Leamer's posterior odds ratio approach. We present results for two alternative specifications: a BVAR with hyperparameters sets using rules of thumb and one with hyperparameters chosen to maximize the predictive density using data from 1980:1 to 1995:4. The prior variance is characterized by a general tightness parameter, a decay parameter and a parameter for lags of other variables. In the first case they are set to 0.2, 1, 0.5, respectively. In the second, they are optimally estimated (point estimates 0.14, 2.06, 1.03). The prior variance on the constant is diffuse. In table 10.1 we report one year ahead Theil-U statistics (the ratio of the MSE of the model to the MSE of a random walk) for the four specifications. Posterior standard error for the two BVAR are in parenthesis.*

Sample	*ARIMA*	*VAR*	*BVAR1*	*BVAR2*
1996:1-2000:4	*1.04*	*1.47*	*1.09 (0.03)*	*0.97 (0.02)*
1990:1-1995:4	*0.99*	*1.24*	*1.04 (0.04)*	*0.94 (0.03)*

Table 10.1: One year ahead Theil-U statistics.

Three features deserve comments. First, forecasting Italian inflation one year ahead is difficult: all models have a hard time to beat a random walk and three of them do worse. Second, an unrestricted VAR performs poorly. Third, a BVAR with default choices is better than a unrestricted VAR but not better than an ARIMA model. Finally, a BVAR with optimally chosen parameters, outperforms both random walk and ARIMA models at the one year horizon but the gains are small. The results are robust: repeating the exercise using data from 1980:1 to 1989:4 to choose the variables, the hyperparameters and estimate the models and data from 1991:1 to 1995:4 to forecast produces qualitatively similar Theil-Us.

10.2.5 Priors derived from DSGE models

The priors we have considered so far are either statistically motivated or based on rules-of-thumb useful for forecasting macroeconomic time series. In both cases, economic theory plays no role, except perhaps in establishing the range of values for the prior distributions. To be able to use BVARs for purposes other than forecasting, one may want to consider priors based on economic theory. In addition, one may be interested in knowing if theory based priors are as good as statistically based priors in forecasting, unconditionally, out-of-sample.

Here we consider priors which are derived from DSGE models. The nature of the model and a prior for the structural parameters imply a prior for the reduced form VAR coefficients. One can dogmatically take these restrictions or simply consider their qualitative content in constructing posterior distributions. In this setup prior information measures the confidence a researcher has that the DSGE structure has generated the observed data.

An alternative representation for the log-linearized solution of a DSGE model is:

$$y_{2t+1} = \mathcal{A}_{22}(\theta)y_{2t} + \mathcal{A}_{23}(\theta)y_{3t+1} \tag{10.15}$$

$$y_{1t} = \mathcal{A}_{12}(\theta)y_{2t} \tag{10.16}$$

where y_{2t} is a $m_2 \times 1$ vector including the states and the driving forces; y_{1t} is $m_1 \times 1$ vecotr including all the endogenous variables and y_{3t+1} are the shocks. Here $\mathcal{A}_{jj'}(\theta)$ are time invariant functions of θ, the vector of structural (preferences, technologies, policy) parameters of the model. It is easy to transform (10.15)-(10.16) into a (restricted) VAR(1) for $y_t = [y_{1t}, y_{2t}]'$ of the form

$$\begin{bmatrix} 0 & 0 \\ 0 & I_{m_2} \end{bmatrix} y_{t+1} = \begin{bmatrix} -I_{m_1} & \mathcal{A}_{12}(\theta) \\ 0 & \mathcal{A}_{22}(\theta) \end{bmatrix} y_t + \begin{bmatrix} 0 \\ \mathcal{A}_{23}(\theta) \end{bmatrix} y_{3t+1} \tag{10.17}$$

or $\mathcal{A}_0 y_{t+1} = \mathcal{A}_1(\theta)y_t + \epsilon_{t+1}(\theta)$ where $\epsilon_{t+1}(\theta) = \begin{bmatrix} 0 \\ \mathcal{A}_{23}(\theta) \end{bmatrix} y_{3t+1}$. Hence, given a prior for θ, the model implies a prior for $\mathcal{A}_{12}(\theta), \mathcal{A}_{22}(\theta), \mathcal{A}_{23}(\theta)$. In turn these priors imply restrictions for the reduced form parameters $A(\ell) = \mathcal{A}_0^{-1}\mathcal{A}_1(\ell)$ and $\Sigma_e = \mathcal{A}_0^{-1}\Sigma_\epsilon \mathcal{A}_0^{-1}$. Expressions for the priors for $\mathcal{A}_{12}(\theta), \mathcal{A}_{22}(\theta), \mathcal{A}_{23}(\theta)$ can be obtained using δ-approximations, i.e. if $\theta \sim \mathbb{N}(\bar{\theta}, \bar{\Sigma}_\theta)$, $vec(\mathcal{A}_{12}(\theta)) \sim \mathbb{N}(vec(\mathcal{A}_{12}(\bar{\theta})), \frac{\partial vec(\mathcal{A}_{12}(\theta))}{\partial \theta} \bar{\Sigma}_\theta \frac{\partial vec(\mathcal{A}_{12}(\theta))}{\partial \theta}')$, etc.

Example 10.6 *Consider a VAR(q): $y_{t+1} = A(\ell)y_t + e_t$. From (10.17) the prior for A_1 is Normal with mean $\mathcal{A}_0^G\mathcal{A}_1(\bar{\theta})$, where $\mathcal{A}_0^G$ is the generalized inverse of $\mathcal{A}_0$ and variance equal to $\Sigma_a = (A_0^G \otimes I_{m_1+m_2})\Sigma_{a_1}(A_0^G \otimes I_{m_1+m_2})'$); where Σ_{a_1} is the variance of $vec(\mathcal{A}_1(\theta))$. A DSGE prior for $A_2, A_3, \ldots$ has a dogmatic form: mean zero and zero variance.*

Since the states of a DSGE model typically include unobservable variables (e.g. the Lagrangian multiplier or the driving forces of the model) or variables measured with error (e.g. the capital stock), it may be more convenient to set up prior restrictions for a VAR composed only of the endogenous variables, as the next example shows.

Example 10.7 *(Ingram and Whiteman). A RBC model with utility function $u(c_t, c_{t-1}, N_t, N_{t-1}) = \ln(c_t) + \ln(1 - N_t)$ implies a law of motion for the states of the form*

$$\begin{bmatrix} K_{t+1} \\ \ln\zeta_{t+1} \end{bmatrix} = \begin{bmatrix} \mathcal{A}_{kk}(\theta) & \mathcal{A}_{k\zeta}(\theta) \\ 0 & \rho_\zeta \end{bmatrix} \begin{bmatrix} K_t \\ \ln\zeta_t \end{bmatrix} + \begin{bmatrix} 0 \\ \epsilon_{1t+1} \end{bmatrix} \equiv \mathcal{A}_{22}(\theta) \begin{bmatrix} K_t \\ \ln\zeta_t \end{bmatrix} + \epsilon_{t+1} \quad (10.18)$$

where K_t is the capital stock and ζ_t is a technological disturbance. The equilibrium mapping between the endogenous variables and the states is $[c_t, N_t, gdp_t, inv_t]' = \mathcal{A}_{12}(\theta) \begin{bmatrix} K_t \\ \ln\zeta_t \end{bmatrix}$ where c_t is consumption, N_t hours, gdp_t output and inv_t investments. Here $\mathcal{A}_{12}(\theta)$ and $\mathcal{A}_{22}(\theta)$ are function of η, the share of labor in production, β the discount factor, δ the depreciation rate, ρ_ζ the AR parameter of the technology shock. Let $y_{1t} = [c_t, N_t, gdp_t, inv_t]'$ and $y_{2t} = [k_t, \ln\zeta_t]'$, $\theta = (\eta, \beta, \delta, \rho_\zeta)$. Then $y_{1t} = A(\theta)y_{1t-1} + e_{1t}$, where $A(\theta) = \mathcal{A}_{12}(\theta)\mathcal{A}_{22}(\theta)(\mathcal{A}_{12}(\theta)'\mathcal{A}_{12}(\theta))^{-1}\mathcal{A}_{12}(\theta)$, $e_{1t} = \mathcal{A}_{12}(\theta)\epsilon_t$ and $(\mathcal{A}_{12}(\theta)'\mathcal{A}_{12}(\theta))^{-1}\mathcal{A}_{12}(\theta)$ is the generalized inverse of $\mathcal{A}_{12}(\theta)$. If $g(\theta)$ is $\theta \sim \mathbb{N}(\begin{bmatrix} 0.58 \\ 0.988 \\ 0.025 \\ 0.95 \end{bmatrix}, \begin{bmatrix} 0.0006 & & & \\ & 0.0005 & & \\ & & 0.0004 & \\ & & & 0.00015 \end{bmatrix})$, the prior mean of $A(\theta)$ is $A(\bar{\theta}) = \begin{bmatrix} 0.19 & 0.33 & 0.13 & -0.02 \\ 0.45 & 0.67 & 0.29 & -0.10 \\ 0.49 & 1.32 & 0.40 & 0.17 \\ 1.35 & 4.00 & 1.18 & 0.64 \end{bmatrix}$ which implies, e.g., substantial feedback from consumption, output and hours to investment (see the last row). The prior variance for $A(\theta)$ is $\Sigma_A = \frac{\partial A(\theta)}{\partial \theta'}\bar{\Sigma}_\theta\frac{\partial A(\theta)}{\partial \theta'}$, where $\frac{\partial A(\theta)}{\partial \theta'}$ is a 16×4 vector. Hence, a RBC prior for y_{1t} implies a normal prior on the first lag with mean $A(\bar{\theta})$ and variance proportional to Σ_A. To relax the dogmatic prior restriction on higher lags, we could assume a Normal prior with zero mean and variance $\propto \frac{\Sigma_A}{h(\ell)}$ where $h(\ell)$ is a decaying function of ℓ.

Exercise 10.11 *(RBC cointegrating prior). In example 10.7 suppose that $(\ln\zeta_t)$ has a unit root. Then all endogenous variables must have a unit root and the stochastic trend is a common one.*
(i) Argue that $(I - \mathcal{A}_{kk}(\theta), -\mathcal{A}_{k\zeta}(\theta))$ must be a cointegrating vector for k_t.
(ii) Argue that $(I_4, -\mathcal{A}_{12}(\theta))$ must be a cointegrating vector for y_{1t}
(iii) Given a Normal prior on θ, derive a cointegrating prior for the $\mathcal{A}'s$.

Exercise 10.12 *Suppose consumers maximize* $u(c_t, c_{t-1}, N_t) = \ln c_t - \epsilon_{2t} \ln N_t$ *subject to the constraint* $c_t + B_{t+1} \leq y_t + (1 + r_t^B)B_t - T_t$ *where* $y_t = N_t \epsilon_{1t}$, ϵ_{1t} *is a technology shock with mean* $\bar{\epsilon}_1$ *and variance* $\sigma^2_{\epsilon_1}$ *and* ϵ_{2t} *is a labor supply shock with mean of* $\bar{\epsilon}_2$ *and variance* $\sigma^2_{\epsilon_2}$. *Here* T_t *are lump sum taxes,* B_t *are real bonds and the government finances a random stream of expenditure using lump sum taxes and real bonds according to the budget constraint* $G_t - T_t = B_{t+1} - (1 + r_t^B)B_t$. *In this model there are three shocks: two supply type shocks* $(\epsilon_{1t}, \epsilon_{2t})$ *and one demand type shock* (G_t).
i) Find a log-linearized solution for N_t, y_t, c_t *and labor productivity* (np_t).
ii) Use the results in i) to construct a prior for a bivariate VAR in hours and output. Derive the posterior distribution for the VAR parameters and the covariance matrix of the shocks. Be precise about the assumptions and the choices you make (Careful, there are three shocks and two variables). Would it make a difference for the answer if you would have used a trivariate model with consumption or labor productivity?
iii) Describe how to construct impulse responses to G_t *shocks using posterior estimates.*
iv) Suppose that, for identification purposes, an investigator makes the assumption that demand shocks have zero contemporaneous effect on hours. Is this assumption reasonable in the logic of the model? Under what conditions the estimated demand shocks you recover from posterior analysis correctly represent G_t *shocks?*

Del Negro and Schorfheide (2003) have suggested an alternative way to append priors derived from DSGE models onto a VAR. The advantage of their approach is that the posterior distributions for both VAR and DSGE parameters can be simultaneously obtained. The basic specification they use differs from the one so far described in an important way. Up to now a DSGE model has provided only the "form" of the prior restrictions (zero mean on lags greater than one, etc.). Here the prior is more tightly based on the data produced by the DSGE model.

The logic of the approach is simple. Since the prior can be thought as an additional observation tagged on to the VAR, one way to add DSGE information is to augment the VAR for the actual data with a prior based on data simulated from the model. The proportion of actual and simulated data points then reflects the relative importance that a researcher gives to the two types of information.

Let the data be represented by a VAR with parameters (α, Σ_e). Assume that $g(\alpha, \Sigma_e)$ is of the form $\alpha \sim \mathbb{N}(\bar{\alpha}(\theta), \bar{\Sigma}(\theta))$; $\Sigma_e^{-1} \sim \mathbb{W}(T_s \bar{\Sigma}_e(\theta), T_s - k)$ where

$$\begin{aligned} \bar{\alpha}(\theta) &= ((X^s)'X^s)^{-1}((X^s)'y^s) \\ \bar{\Sigma}(\theta) &= \Sigma_e(\theta) \otimes ((X^s)'X^s)^{-1} \\ \bar{\Sigma}_e(\theta) &= (y^s - X^s\bar{\alpha}(\theta))(y^s - (X^s)\bar{\alpha}(\theta))' \end{aligned} \tag{10.19}$$

Here y^s is data simulated from the DSGE model, $X^s = (I_m \otimes \mathsf{X}^s)$ is a matrix of lags in the VAR representation of simulated data and θ the structural parameters. In (10.19), the moments of $g(\alpha, \Sigma_e)$ depend on θ through the simulated data (y^s, X^s). If T_s measures the length of simulated data, $\kappa = \frac{T_s}{T}$ controls the relative importance of the information

contained in actual and simulated data. Clearly, if $\kappa \to 0$, the actual data dominates and if $\kappa \to \infty$, the simulated data dominates.

The model has a hierarchical structure $f(\alpha, \Sigma_e|y)g(\alpha|\theta)g(\Sigma_e|\theta)g(\theta)$. Conditional on θ, the posterior for α, Σ_e are easily derived. In fact, since the likelihood and the prior are conjugate $(\alpha|\theta, y, \Sigma_e) \sim \mathbb{N}(\tilde{\alpha}(\theta), \tilde{\Sigma}(\theta))$; $(\Sigma_e^{-1}|\theta, y) \sim \mathbb{W}((\kappa+T)\tilde{\Sigma}_e(\theta), T+\kappa-k)$ where

$$\begin{aligned}
\tilde{\alpha}(\theta) &= (\kappa\frac{(X^s)'X^s}{T^s} + \frac{X'X}{T})^{-1}(\kappa\frac{(X^s)'y^s}{T^s} + \frac{X'y}{T}) \\
\tilde{\Sigma}(\theta) &= \Sigma_e(\theta) \otimes ((X^s)'X^s + X'X)^{-1} \qquad (10.20) \\
\tilde{\Sigma}_e(\theta) &= \frac{1}{(1+\kappa)T}[(y^s)'y^s + y'y) - ((y^s)'X^s + y'X)((X^s)'X^s + X'X)^{-1}((X^s)'y^s + X'y)]
\end{aligned}$$

where $X = (I \otimes \mathsf{X})$. The posterior for θ can be computed using the hierarchical structure of the model. In fact, $g(\theta|y) \propto f(\alpha, \Sigma_e, y|\theta)g(\theta)$ where $f(\alpha, \Sigma_e, y|\theta) \propto |\Sigma_e|^{-0.5(T-m-1)} \exp\{-0.5 tr[\Sigma_e^{-1}(y-X\alpha)'(y-X\alpha)\} \times |\bar{\Sigma}_e(\theta)|^{-0.5(T_s-m-1)} \exp\{-0.5tr[\Sigma_e^{-1}(y^s - X^s\bar{\alpha}(\theta))'(y^s - X^s\bar{\alpha}(\theta))\}$. We will discuss how to draw from this posterior in chapter 11.

Exercise 10.13 *Use the fact that $g(\alpha, \Sigma_e, \theta|y) = g(\alpha, \Sigma_e|y, \theta)g(\theta|y)$, to suggest an algorithm to draw sequences for (α, Σ_e). How do you compute impulse responses in the VAR?*

Exercise 10.14 *Suppose $g(\Sigma_e)$ is non-informative. Show the form of $(\tilde{\alpha}, \tilde{\Sigma}_e)$ in this case.*

All posterior moments in (10.20) are conditional on a value of κ. Since this parameter regulates the relative importance of sample and prior information it is important to appropriately select it. As in standard BVAR, there are two ways to proceed. First, we can use a rule of thumb, e.g. set $\kappa = 1$, meaning that T simulated data are added to the actual ones. Second, we can choose it to maximize the predictive density of the model.

Exercise 10.15 *Show the form of $f(y|\kappa)$. Describe how to find its maximum numerically.*

Exercise 10.16 *Consider the working capital model described in exercise 1.14 of chapter 2 driven by shocks to technology, government expenditure and the monetary policy rule. Choose appropriate priors for the parameters (for example, Normal, Gamma or Beta for parameters that lie in an interval). Simulate data for output, inflation and the nominal interest rate. Combine this data with actual data for output, inflation and the nominal interest rate. Explore the predictive density of inflation numerically for different values of κ. Is there a relationship between the κ which maximizes the predictive density and the one which minimizes the MSE of the forecasts? How would you compare such a model against a sticky price, sticky wage model?*

10.2.6 Probability distributions for forecasts: Fan Charts

BVAR models can be used to construct probability distributions for future events and therefore are well suited to produce e.g. fan charts or probabilities of turning points. To see how this can be done, set $\bar{y} = 0$ and rewrite the VAR model in a companion form

$$\mathbb{Y}_t = \mathbb{A}\mathbb{Y}_{t-1} + \mathbb{E}_t \tag{10.21}$$

where $\mathbb{Y}_t$ and $\mathbb{E}_t$ are $mq \times 1$ vectors, $\mathbb{A}$ is a $mq \times mq$ matrix.

Repeatedly substituting we have $\mathbb{Y}_t = \mathbb{A}^\tau \mathbb{Y}_{t-\tau} + \sum_{j=0}^{\tau-1} \mathbb{A}^j \mathbb{E}_{t-j}$ or $y_t = \mathbb{S}\mathbb{A}^\tau \mathbb{Y}_{t-\tau} + \sum_{j=0}^{\tau-1} \mathbb{A}^j e_{t-j}$ where $\mathbb{S}$ is such that $\mathbb{S}\mathbb{Y}_t = y_t$, $\mathbb{S}\mathbb{E}_t = e_t$ and $\mathbb{S}'\mathbb{S}E_t = E_t$. A "point" forecast for $y_{t+\tau}$ is obtained plugging-in some location measures of the posterior of $\mathbb{A}$ into $y_t(\tau) = \mathbb{S}\mathbb{A}^\tau \mathbb{Y}_t$. Call this point forecast $\hat{y}_t(\tau)$. The forecast error is $y_{t+\tau} - \hat{y}_t(\tau) = \sum_{j=0}^{\tau-1} \mathbb{A}^j e_{t+\tau-j} + [y_t(\tau) - \hat{y}_t(\tau)]$ and the variance of the forecast error can be computed once posterior estimates of $\mathbb{A}$ are available. This is easy when $\tau = 1$. For $\tau \geq 2$ only approximate expressions for the MSE are available (see e.g. Lutkepohl (1991), p. 88).

Exercise 10.17 *Show the MSE of the forecasts when $\tau = 1$.*

When a distribution of forecasts is actually needed we can exploit the fact that we can draw from $g(\alpha|y)$. We describe how "fan charts" can be obtained for case 1. prior with the obvious extension if also Σ_e is a random variable. Let $\tilde{\mathcal{P}}\tilde{\mathcal{P}}'$ be any orthogonal factorization of Σ_e. Then, at a given t:

Algorithm 10.1

1) *Draw v_a^l from a $\mathbb{N}(0,1)$ and set $\alpha^l = \tilde{\alpha} + \tilde{\mathcal{P}}^{-1} v_a^l$, $l = 1, \ldots L$.*

2) *Construct point forecasts $y_t^l(\tau)$, $\tau = 1, 2, , \ldots$ conditioning on α^l.*

3) *Construct distributions at each τ using kernel methods and extract percentiles.*

Exercise 10.18 *Consider case 4. prior (i.e. a Normal prior for α and a Wishart prior for Σ_e^{-1}). Modify algorithm 10.1 to fit this situation.*

Algorithm 10.1 can also be used recursively, using estimates of $\tilde{\alpha}$ which are updated through the sample. The only difference is that $\tilde{\alpha}$ and $\tilde{\mathcal{P}}$ now depend on t.

Example 10.8 *In certain situations one wants to compute "average" forecasts at step τ, i.e. may want to compute the predictive density $f(y_{t+\tau} \mid y_t) = \int f(y_{t+\tau} \mid y_t, \alpha) g(\alpha \mid y_t) d\alpha$ where $f(y_{t+\tau} \mid y_t, \alpha)$ is the conditional density of the future observation vector, given α and the model, and $g(\alpha \mid y_t)$ is the posterior of α at t. Given draws from algorithm 10.1 and the model then $\hat{y}_t(\tau) = L^{-1} \sum_{l=1}^{L} y_t^l(\tau)$ and its numerical variance is $L^{-1} \sum_{l=1}^{L} \sum_{j=-J(L)}^{J(L)} \mathcal{K}(j) ACF_\tau^l(j)$, where $\mathcal{K}(j)$ is a kernel and $ACF_\tau(j)$ the autocovariance of $\hat{y}_t(\tau)$ at lag j.*

Turning point probabilities can also be computed from the numerically constructed predictive density of future observations. For example, given $y_t^l(\tau),\ l = 1, \dots L$ we only need to check if e.g. a two quarters rule is satisfied for each draw α^l. The fraction of draws for which the condition is satisfied is an estimate of the probability of the event at $t + \tau$.

Example 10.9 *Continuing with example 10.5, figure 10.2 presents BVAR based 68 and 95 percent bands for inflation forecasts one year ahead where we recursively update posterior estimates. The forecasting sample is 1996:1-1998:2. The bands are relatively tight reflecting very precise estimates. This precision can also be seen from the distribution of the forecasts one year ahead, constructed with data up to 1995:4. We calculate the distribution of the number of downturns that the annualized inflation rate is expected to experience over the sample 1996:1-2000:4. Downturns are identified with a two quarters rule. In the actual data there are four downturns. The median number of forecasted downturns is three. Moreover, in 90 per cent of the cases the model underpredicts the actual number of downturns and it never produces more than four downturns.*

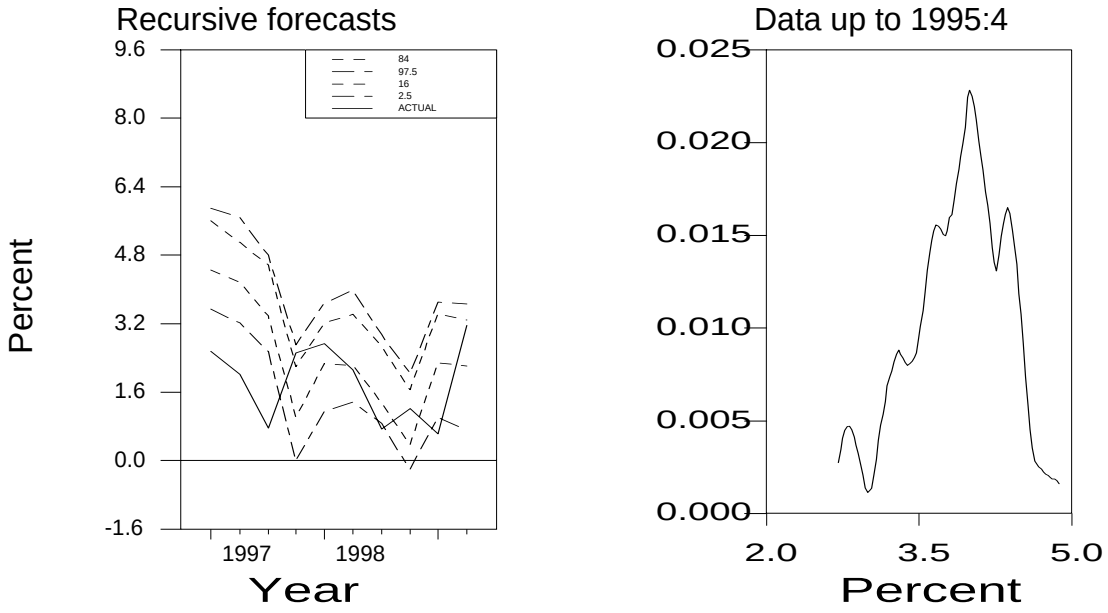


Figure 10.2: Forecasts of Italian inflation.

10.3 Structural BVARs

The priors we have specified in section 10.2 are designed for reduced form VAR models. What kind of priors are reasonable for structural VARs?

There are two approaches in the literature. A naive one, employed by Canova (1991), Gordon and Leeper (1994), is to use a Normal-Wishart structure for reduced form parameters (α, Σ_e). Then draws for the structural parameters are made conditional on the identification restrictions. Hence, if $\Sigma_e = \mathcal{A}_0^{-1}\mathcal{A}_0^{-1'}$, then $A_j = \mathcal{A}_0^{-1}\mathcal{A}_j$, where A_j are VAR coefficients. This approach is appropriate if $\mathcal{A}_0$ is just identified since there is a unique mapping between draws of Σ_e and draws of $\mathcal{A}_0$. When $\mathcal{A}_0$ is overidentified this method neglects the (over-identifying) restrictions. In this case, it is better to work with the structural model, and the prior suggested by Sims and Zha (1998). Consider the following structural model, where $\mathcal{A}_0$ is non singular and $\bar{y}$ only includes deterministic variables:

$$\mathcal{A}_0 y_t - \mathcal{A}(\ell) y_{t-1} + \mathcal{C}\bar{y}_t = \epsilon_t \quad \epsilon_t \sim (0, I) \tag{10.22}$$

where $\mathcal{A}(\ell) = \mathcal{A}_1\ell + \ldots \mathcal{A}_q\ell^q$. Staking the t observations we have:

$$\mathsf{Y}\mathcal{A}_0 - \mathsf{X}\mathcal{A}_- = \varepsilon \tag{10.23}$$

where Y is a $T \times m$, X is a $T \times k$ matrix of lagged and exogenous variables, $k = mq + m_c$; ε is a $T \times m$ matrix. Let $\mathsf{Z} = [\mathsf{Y}, -\mathsf{X}]$; $\mathcal{A} = [\mathcal{A}_0, \mathcal{A}_-]'$. The likelihood function is:

$$L(\mathcal{A}|y) \propto |\mathcal{A}_0|^T \exp\{-0.5tr(\mathsf{Z}\mathcal{A})'(\mathsf{Z}\mathcal{A})\} = |\mathcal{A}_0|^T \exp\{-0.5b'(I_{mk} \otimes \mathsf{Z}'\mathsf{Z})b\} \tag{10.24}$$

where $b = vec(\mathcal{A})$ is a $m(k+m) \times 1$ vector; $b_0 = vec(\mathcal{A}_0)$ is a $m^2 \times 1$ vector; $b_- = vec(\mathcal{A}_-)$ is a $mk \times 1$ vector and I_{mk} is a $(mk \times mk)$ matrix.

Suppose $g(b) = g(b_0)g(b_-|b_0)$ where $g(b_0)$ may have singularities (due to zero identification restrictions) and let $g(b_-|b_0) \sim \mathbb{N}(\bar{h}(b_0), \bar{\Sigma}(b_0))$. The posterior is :

$$g(b|y) \propto g(b_0)|\mathcal{A}_0|^T|\Sigma(b_0)|^{-0.5} \exp\{-0.5[b'(I_{mk}\otimes\mathsf{Z}'\mathsf{Z})b\} \exp\{(b_- - \bar{h}(b_0))'\bar{\Sigma}(b_0)^{-1}(b_- - \bar{h}(b_0))\} \tag{10.25}$$

Since $b'(I_{mk}\otimes\mathsf{Z}'\mathsf{Z})b = b_0'(I_{mk}\otimes\mathsf{Y}'\mathsf{Y})b_0 + b_-'(I_{nk}\otimes\mathsf{X}'\mathsf{X})b_- - 2b_-'(I_{mk}\otimes\mathsf{X}'\mathsf{Y})b_0$, conditional on b_0, the quantity in the exponent in (10.25) is quadratic in b_- so that $g(b_-|b_0, y) \sim \mathbb{N}(\tilde{h}(b_0), \tilde{\Sigma}(b_0))$ where $\tilde{h}(b_0) = \tilde{\Sigma}(b_0)((I_{mk} \otimes \mathsf{X}'\mathsf{Y})\hat{h}(b_0) + \bar{\Sigma}(b_0)^{-1}\bar{h}(b_0))$ and $\tilde{\Sigma}(b_0) = ((I \otimes \mathsf{X}'\mathsf{X}) + \bar{\Sigma}(a_0)^{-1})^{-1}$. Furthermore

$$\begin{aligned} g(b_0|y) \propto\ & g(b_0)|\mathcal{A}_0|^T|(I_{mk} \otimes \mathsf{X}'\mathsf{X})\bar{\Sigma}(b_0) + I|^{-0.5} \\ & exp\{-0.5[b_0'(I_{mk} \otimes \mathsf{Y}'\mathsf{Y})b_0 + h(b_0)'\bar{\Sigma}(b_0)^{-1}h(b_0) - \tilde{h}(b_0)\tilde{\Sigma}(b_0)\tilde{h}(b_0)]\} \end{aligned} \tag{10.26}$$

Since $dim(b_-) = mk$, the calculation of $g(b_-|b_0, y)$ may be time consuming. Equation by equation computations are possible if the structural model is in SUR format, i.e. if we can run m separate least square regressions with k parameters each. To do this we need to choose $\bar{\Sigma}(b_0)$ appropriately. For example, if $\bar{\Sigma}(b_0) = \bar{\Sigma}_1 \otimes \bar{\Sigma}_2$ and $\bar{\Sigma}_1 \propto I$, then even if $\bar{\Sigma}_{2i} \neq \bar{\Sigma}_{2j}$, independence across equations is guaranteed since $(I_{mk} \otimes \mathsf{X}'\mathsf{X}) + \bar{\Sigma}(b_0)^{-1} \propto (I_{mk} \otimes \mathsf{X}'\mathsf{X}) + \text{diag}\{\bar{\Sigma}_{21}, \ldots, \bar{\Sigma}_{2m}\} = \text{diag}\{\bar{\Sigma}_{21} + \mathsf{X}'\mathsf{X}, \ldots, \bar{\Sigma}_{2m} + \mathsf{X}'\mathsf{X}\}$.

Note that if we had started from a reduced form VAR (as we have done in exercise 10.9) the structure of $\tilde{\Sigma}(b_0)$ would have been $\tilde{\Sigma}(b_0) = [(\Sigma_e \otimes \mathsf{X}'\mathsf{X}) + \bar{\Sigma}(b_0)^{-1}]^{-1}$, where Σ_e is

the covariance matrix of the disturbances. This means that to maintain the computations simple $\bar{\Sigma}(b_0)$ must allow correlation across equations (contrary, for example, to what the Minnesota prior assumes).

It is interesting to map structural priors into Minnesota priors. Let $\mathcal{A}_0$ be given and let the VAR be $y_t = A(\ell)y_{t-1} + C\bar{y}_t + e_t$. Let $\alpha = vec[A_1, \dots A_q, C]$. Since $A(\ell) = [\mathcal{A}_-\mathcal{A}_0^{-1}]$; $E(\alpha) = [I_m, 0, \dots 0]$ and $var(\alpha) = \bar{\Sigma}_\alpha$ where $\bar{\Sigma}_\alpha$ was defined in (10.12) imply

$$E(\mathcal{A}_-|\mathcal{A}_0) = [\mathcal{A}_0, 0, \dots, 0] \tag{10.27}$$

$$var(\mathcal{A}_-|\mathcal{A}_0) = \text{diag}(b_{-(ijl)}) = \frac{\phi_0\phi_1}{h(\ell)\sigma_j^2} \quad i, j = 1, \dots m,\ \ell = 1, \dots, q \tag{10.28}$$

$$= \phi_0\phi_2 \qquad \text{otherwise} \tag{10.29}$$

where i stands for equation, j for variable, ℓ for lag, ϕ_0 (ϕ_1) controls the tightness of the prior variance of $\mathcal{A}_0$, ($\mathcal{A}_+$) and ϕ_2 the tightness of the prior variance of $\mathcal{C}$.

Three features of (10.27)-(10.29) are worth mentioning: (i) there is no distinction between own and other coefficients since, in simultaneous equation models, no normalization with one right hand side variable is available; (ii) the scale factors differ from those of reduced form BVARs since $var(\epsilon_t) = I$; (iii) since $\alpha = vec[\mathcal{A}_+\mathcal{A}_0^{-1}]$ beliefs about α may be correlated across equations (if beliefs about $\mathcal{A}_0$ are).

As in a reduced form BVARs, stochastic linear restrictions can be added to the specification and combined with the data using the logic of Theil's mixed estimation.

Exercise 10.19 *(Controlling for trends: sum of coefficients restrictions) Suppose the average value of lagged y_i's (say, $\bar{\bar{y}}_i$) is a good predictor of y_{it} for equation i. Write this information as $\mathsf{Y}^\dagger\mathcal{A}_0 - \mathsf{X}^\dagger\mathcal{A}_- = V$ where $y^\dagger = \{y_{ij}^\dagger\} = \phi_3\bar{\bar{y}}_i$ if $i = j$ and zero otherwise, $i, j = 1, \dots m$; $x^\dagger = \{x_{i\tau}^\dagger\} = \phi_3\bar{\bar{y}}_i$ if $i = j$, for $\tau < k$ and zero otherwise, $i = 1, \dots m,\ \tau = 1, \dots k$. Construct the posterior for b_- under this restriction.*

Adding the sum of coefficient restrictions introduces correlation among the coefficients of a variable in an equation. When $\phi_3 \to \infty$, the restriction implies a model in first difference, i.e. the model has m unit roots and no cointegration.

Exercise 10.20 *(Controlling for seasonality: seasonal sum of coefficients restrictions). Suppose the average value of y_{t-j} is good predictor of y_t for each equation. Setup this restriction as a dummy observation and construct the posterior for b_-.*

Exercise 10.21 *(Controlling for cointegration: initial dummy restriction) Suppose we set up an initial dummy observation of the form $\mathsf{Y}^\ddagger\mathcal{A}_0 - \mathsf{X}^\ddagger\mathcal{A}_- = V$ where $y^\ddagger = \{y_j^\ddagger\} = \phi_4\bar{\bar{y}}_j$ if $j = 1, \dots M$, $x^\ddagger = \{x_\tau^\ddagger\} = \phi_4\bar{\bar{y}}_j$ if $\tau \leq k-1$ and $X^\ddagger = \phi_4$ if $\tau = k$. Construct the posterior for b_- under this additional restriction.*

The prior of exercise 10.21 forces all the variables to be stationary. In fact, if $\phi_4 \to \infty$, the dummy observation becomes $[I - \mathcal{A}_0^{-1}\mathcal{A}(1)]\bar{y}_0 + \mathcal{A}_0^{-1}\mathcal{C} = 0$. If $\mathcal{C} = 0$, there is a one unit root, while if $\mathcal{C} \neq 0$ there are no unit roots.

To calculate (10.26) we need $g(b_0)$. Since for identification purposes, some elements of b_0 may be forced to be zero, we make a distinction between hard restrictions (those imposing identification, possibly of blocks of equations) and soft restrictions (those involving a prior on non-zero coefficients). Since little is typically known about b_0, a non-informative prior should be preferred i.e. $g(b_0^0) \propto 1$ where b_0^0 are the non-zero elements of b_0. In some occasions, a Normal prior may also be appropriate.

Example 10.10 *Suppose we have $m(m-1)/2$ restrictions so that $\mathcal{A}_0$ is just identified. Assume, for example, that $\mathcal{A}_0$ is lower triangular and let b_0^0 be the nonzero elements of $\mathcal{A}_0$. Suppose $g(b_0^0) = \prod_i g(b_{0i}^0)$, where each $g(b_{0i}^0)$ is $\mathbb{N}(0, \sigma^2(b_0^0))$ so that the coefficients of, say, GDP and unemployment in the first equation may be related to each other but are unrelated with the coefficients of GDP and unemployment in other equations. Set, for example, $\sigma^2(b_{0ij}^0) = (\frac{\phi_5}{\sigma_i})^2$ i.e. all the elements of equation i have the same variance. Since the system is just identified one can also use a Wishart prior for Σ_e^{-1}, with $\bar{\nu}$ degrees of freedom and scale matrix $\bar{\Sigma}$ to derive a prior for b_0^0. Since a lower triangular $\mathcal{A}_0$ is just the Choleski factor of Σ_e^{-1}, if $\bar{\nu} = m+1$, $\bar{\Sigma} = \text{diag}\ (\frac{\phi_5}{\sigma_i})^2$, then a prior for b_0^0 is proportional to $\mathbb{N}(0, \sigma^2(b_0^0))$, where the factor of proportionality is the Jacobian of the transformation, i.e. $|\frac{\partial \Sigma_e^{-1}}{\partial \mathcal{A}_0}| = 2^m \prod_{j=1}^m b_{jj}^j$. Since the likelihood contains a term $|\mathcal{A}_0|^T = \prod_{j=1}^T b_{jj}^T$, ignoring the Jacobian is irrelevant if $T >> m$.*

The posterior $g(b_0|y)$ can not be computed analytically. To simulate a sequence we can use one of the algorithms we described in chapter 9. For example, one could:

Algorithm 10.2

1) *Calculated posterior mode b_0^* of $g(b_0|y)$ and the Hessian at b_0^*.*

2) *Draw b_0 from a normal centered at b_0^* with covariance equal to the Hessian at b_0^* or a t-distribution with the same mean and covariance and $\nu = m+1$ degrees of freedom.*

3) *Use importance sampling to weight the draws, checking the magnitude of $IR^l = \frac{\tilde{g}(b_0^l)}{g^{IS}(b_0^l)}$, where $g^{IS}(b_0)$ is an importance density, and $l = 1, \ldots, L$.*

As alternative one could use a Metropolis-Hastings (MH) algorithm with a Normal or a t-distribution as the target, or the restricted Gibbs sampler of Waggoner and Zha (2003).

Exercise 10.22 *Describe how to use a MH algorithm to draw a sequence from $g(b_0|y)$.*

It is immediate to extend the framework to the case where non-contemporaneous restrictions are used to identify the VAR.

Exercise 10.23 *Suppose $\mathcal{A}_0$ is just identified using long run restrictions. How would you modify the prior for $\mathcal{A}_0$ to account for this?*

Exercise 10.24 *Suppose $\mathcal{A}_0$ is overidentified. How should the prior for $\mathcal{A}_0$ be changed?*

Exercise 10.25 *Suppose $\mathcal{A}_0$ is identified using sign restrictions. Let $\Sigma_e = \tilde{\mathcal{P}}(\omega)\tilde{\mathcal{P}}'(\omega)$, where ω is an angle. How would you modify the prior for $\mathcal{A}_0$ to take this into account? How would you modify the algorithm to draw from the posterior distribution of $\mathcal{A}_0$? (Hint: treat ω as a random variable and select an appropriate prior distribution)*

There are a number of extensions one can consider. Here we analyze two:

1. Structural VAR models with exogenous stochastic variables: e.g. oil prices in a structural VAR for domestic variables.

2. Structural VAR models with block exogenous variables and overidentifying restrictions in some block, e.g. a two-country structural model where one is block exogenous.

We assume that y_t is demeaned so that $\bar{y}_t$ is omitted from the model. For the case of structural models with exogenous variables, let

$$\mathcal{A}_{i0}y_t - \mathcal{A}_i(\ell)y_{t-1} = \epsilon_{it} \quad \epsilon_{it} \sim \mathbb{N}(0, I) \tag{10.30}$$

where $i = 1, \ldots, n$ refers to the number of blocks; $m = \sum_{i=1}^{n} m_i$ with m_i equations in each block; ϵ_{it} is $m_i \times 1$ for each i, $\mathcal{A}_i(\ell) = (\mathcal{A}_{i1}(\ell), \ldots, \mathcal{A}_{in}(\ell))$ and each $\mathcal{A}_{ij}(\ell)$ is a $m_i \times m_j$ matrix for each ℓ. (10.30) is just the block representation of (10.22). Rewrite (10.30) as

$$y_{it} = A_i(\ell)y_{it-1} + e_{it} \tag{10.31}$$

where $A_i(\ell) = (0_{i-}, I_i, 0_{i+}) - \mathcal{A}_{i0}^{-1}\mathcal{A}_i(\ell)$; 0_{i-} is a matrix of zeros of dimension $m_i \times m_{i_-}$, 0_{i+} is a matrix of zeros of dimension $m_i \times m_{i_+}$, where $m_{i_-} = 0$ for $i = 1$ and $m_{i_-} = \sum_{j}^{i-1} m_j$ for $i = 2, \ldots, n$; $m_{i_+} = 0$ for $i = n$ and $m_{i_+} = \sum_{j=i+1}^{n} m_j$ for $i = 1, \ldots, n-1$ and where $E(e_t e_t') = \text{diag}\{\Sigma_{ii}\} = \text{diag}\{\mathcal{A}_{i0}^{-1}\mathcal{A}_{i0}^{-1\prime}\}$. Stacking the T observations to have

$$\mathsf{Y}_i = \mathsf{X}_i\mathsf{A}_i + \mathsf{E}_i \tag{10.32}$$

where Y_i and E_i are $T \times m_i$ matrices, X_i is a $T \times k_i$ matrix and k_i is the number of coefficients in each block. The likelihood function is

$$\begin{aligned} f(A_i, \Sigma_{ii}|y_T, \ldots, y_1, y_0 \ldots) &\propto \prod_{i=1}^{n} |\mathcal{A}_{i0}|^T \exp\{-0.5tr[(\mathsf{Y}_i - \mathsf{X}_i\mathsf{A}_i)'(\mathsf{Y}_i - \mathsf{X}_i\mathsf{A}_i)\mathcal{A}_{i0}'\mathcal{A}_{i0}]\} \\ &\propto \prod_{i=1}^{n} |\mathcal{A}_{i0}|^T \exp\{-0.5tr[(\mathsf{Y}_i - \mathsf{X}_i\mathsf{A}_{i,ols})'(\mathsf{Y}_i - \mathsf{X}_i\mathsf{A}_{i,ols})\mathcal{A}_{i0}'\mathcal{A}_{i0} \\ &+ (\mathsf{A}_i - \mathsf{A}_{i,ols})'\mathsf{X}_i'\mathsf{X}_i(\mathsf{A}_i - \mathsf{A}_{i,ols})\mathcal{A}_{i0}'\mathcal{A}_{i0}]\} \end{aligned} \tag{10.33}$$

where $\mathsf{A}_{i,ols} = (\mathsf{X}_i'\mathsf{X}_i)^{-1}(\mathsf{X}_i'\mathsf{Y}_i)$ and tr indicates the trace of the matrix. Suppose $g(\mathcal{A}_{i0}, \mathcal{A}_i) \propto |\mathcal{A}_{i0}|^{k_i}$. Then the posterior for $\mathcal{A}_{i0}$ and $\alpha_i = vec(\mathcal{A}_i)$ has the same form as the likelihood and

$$\begin{aligned} g(\mathcal{A}_{i0}|y) &\propto |\mathcal{A}_{i0}|^T \exp\{-0.5tr[(\mathsf{Y}_i - \mathsf{X}_i\mathsf{A}_{i,ols})'(\mathsf{Y}_i - \mathsf{X}_i\mathsf{A}_{i,ols})\mathcal{A}_{i0}'\mathcal{A}_{i0}]\} \quad (10.34) \\ g(\alpha_i|\mathcal{A}_{i0}, y) &\sim \mathbb{N}(\alpha_{i,ols}, (\mathcal{A}_{i0}'\mathcal{A}_{i0})^{-1} \otimes (\mathsf{X}_i'\mathsf{X}_i)^{-1}) \quad (10.35) \end{aligned}$$

where $\alpha_{i,ols} = vec(\mathsf{A}_{i,ols})$. As before, if $\mathcal{A}_{i0}$ is the Choleski factor of Σ_{ii}^{-1} and $g(\Sigma_{ii}^{-1}) \propto |\Sigma_{ii}^{-1}|^{0.5k_i}$, then the posterior for Σ_{ii}^{-1} has Wishart form with parameters $([(\mathsf{Y}_i - \mathsf{X}_i\mathsf{A}_{i,ols})'(\mathsf{Y}_i - \mathsf{X}_i\mathsf{A}_{i,ols})]^{-1}, T - m_i - 1)$. Hence, one could draw from the posterior of Σ_{ii}^{-1} and use the Choleski restrictions to draw $\mathcal{A}_{i0}$. When $\mathcal{A}_{i0}$ is overidentified, we need to draw $\mathcal{A}_{i0}$ from the marginal posterior (10.35), which is of unknown form. To do so one could use, e.g., a version of the importance sampling algorithm 10.2.

Exercise 10.26 *Extend algorithm 10.2 to the case where the VAR has different lags in different blocks.*

Exercise 10.27 *Suppose $g(\mathcal{A}_i) \sim \mathbb{N}(\bar{\mathcal{A}}_i, \bar{\Sigma}_{\mathcal{A}})$. Show the form of $g(\alpha_i|\mathcal{A}_{i0}, y)$ in this case.*

For the case of block exogenous variables with overidentifying restrictions, suppose there are linear restrictions on $\mathcal{A}_{ij0}, \quad j > i$. This case is different from the previous case since overidentifying restrictions were placed on $\mathcal{A}_{ii0}$. Define $\mathcal{A}_i^*(\ell) = \mathcal{A}_{i0} - \mathcal{A}_i(\ell), \ i = 1, \ldots n$ and rewrite the system as $\mathcal{A}_{i0}y_t = \mathcal{A}_i^*(\ell)y_t + \epsilon_{it}$. Stacking the observations we have

$$\mathsf{Y}\mathcal{A}_{i0}' = \mathsf{X}_i\mathbb{A}_i^* + \epsilon_i \quad (10.36)$$

where X_i is a $T \times k_i^*$ matrix including all right hand side variables, $k_i^* = k_i - m_{i+1} - \ldots - m_n$; $\mathbb{A}_i^*$ is a $k_i^* \times m_i$ companion matrix of $\mathcal{A}_i^*(\ell)$; ϵ_i a $T \times m_i$ matrix; $Y = [Y_1, \ldots, Y_n]$ is a $T \times m$ matrix ; $\mathcal{A}_{i0} = \{\mathcal{A}_{i10}, \ldots \mathcal{A}_{in0}; \ \mathcal{A}_{ij0} = 0, j < i\}$ is a $m \times m_i$ matrix. Let $\mathsf{A}_{i,ols}^* = (\mathsf{X}_i'\mathsf{X}_i)^{-1}\mathsf{X}_i'\mathsf{Y}$ and let the prior for $(\mathcal{A}_i(0), \mathsf{A}_i^*)$ be non-informative. Letting $\alpha_i^* = vec(\mathsf{A}_i^*)$, the posteriors are:

$$\begin{aligned} g(\mathcal{A}_{i0}|y) &\propto |\mathcal{A}_{i0}|^T \exp\{-0.5tr[(\mathsf{Y}_i - \mathsf{X}_i\mathsf{A}_{i,ols}^*)'(\mathsf{Y}_i - \mathsf{X}_i\mathsf{A}_{i,ols}^*)\mathcal{A}_{i0}'\mathcal{A}_{i0}]\} \\ g(\alpha_i^*|\mathcal{A}_{i0}, y) &\sim \mathbb{N}(\alpha_{i,ols}^*, (I_i \otimes (\mathsf{X}_i'\mathsf{X}_i)^{-1})) \quad (10.37) \end{aligned}$$

Exercise 10.28 *Describe how to draw posterior sequences for $(\alpha_i^*, \mathcal{A}_{i0})$ from (10.37).*

We conclude with an example illustrating the techniques described in this section.

Example 10.11 *We take monthly US data from 1959:1 to 2003:1 for the log of GDP, the log of CPI, log of M2, the Federal funds rate and log of commodity prices. We are interested in the dynamic responses of the first four variables to an identified monetary policy shock and in knowing how much of the variance of output and inflation is explained by monetary policy shocks. We use contemporaneous restrictions and overidentify the system by assuming*

that the monetary authority only looks at money when manipulating the Federal funds rate. Hence, the system has a Choleski form (in the order in which the variables are listed) except for the (3,1) entry which is set to zero. We assume $b_0^0 \sim \mathbb{N}(0, I)$ *and use as importance sampling a Normal centered at the mode and with dispersion equal to the Hessian at the mode. We monitor the draws using the importance ratio and find that in only 11 out of 1000 draws the weight given to the draw is large.*

The median response and the 68% band for each variable are in figure 10.3. Both output and money persistently decline in response to an interest rate increase. The response of prices is initially zero but turns positive and significant after a few quarters - a reminiscence of what is typically called the "price puzzle". Monetary shocks explain 4-18 per cent of the variance of output at the 20 quarters horizon and only 10-17 per cent of the variance of prices. One may wonder what moves prices then: it turns out that output shocks explain 45-60 per cent of the variability of prices in the sample.

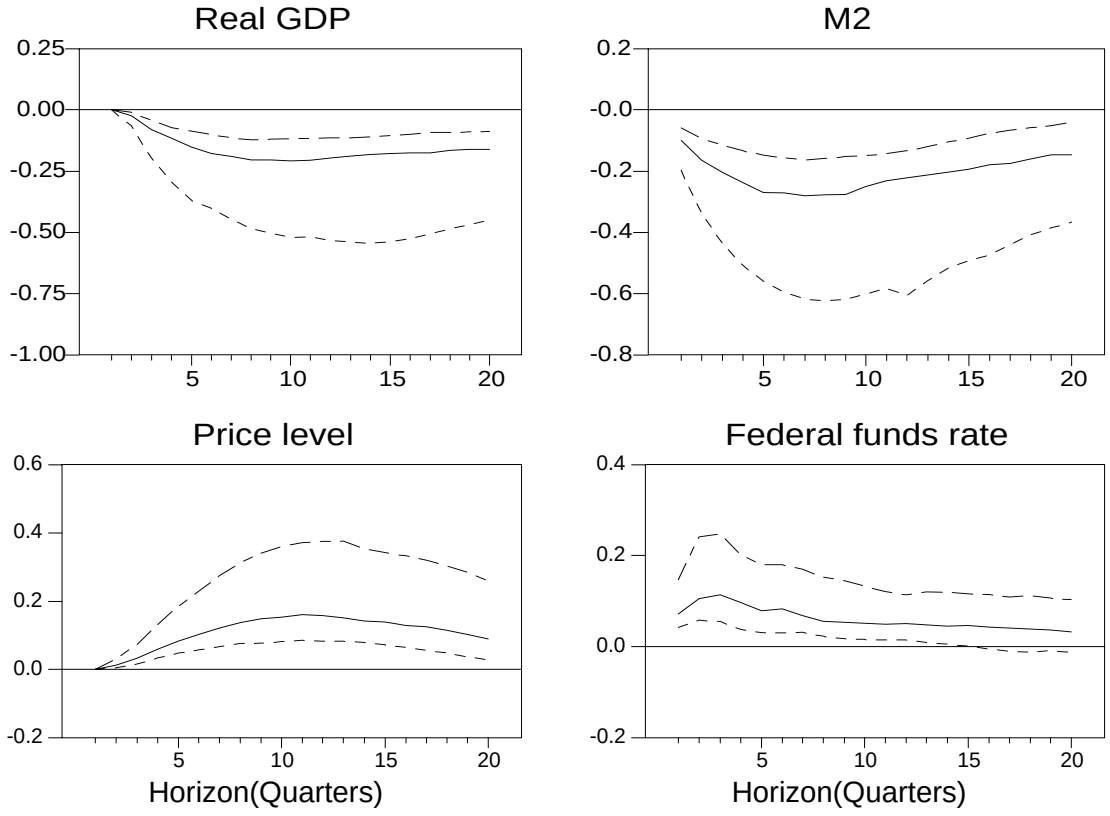


Figure 10.3: Median and 68% band for the responses to a US monetary policy shock.

10.4 Time Varying Coefficients BVARs

Economic time series tend to show evolving features. One could think of these changes as abrupt and model the switch as a structural break (either in the intercept, in the slope coefficients or in both). Alternatively, one may suspect that changes are related to some unobservable state, for example, the business cycle, in which case the coefficients or the covariance matrix or both could be made a function of a finite order Markov Chain (as we will do in Chapter 11). Since structural changes are rare but the coefficients tend evolve continuously one may finally prefer a model with smoothly changing coefficients. Time varying coefficient models have a long history in applied work going back, at least, to Cooley and Prescott (1973), and classical estimation methods, ranging from generalized least square (Swaamy (1970)) to Kalman filtering, are available. Here we treat the law of motion of the coefficients as the first layer of an hierarchical prior and specify, in a second layer, the distributions for the parameters of this law of motion.

The model we consider is of the form

$$y_t = A_t(\ell)y_{t-1} + C_t\bar{y}_t + e_t \qquad e_t \sim \mathbb{N}(0, \Sigma_e) \tag{10.38}$$

$$\alpha_t = \mathbb{D}_1\alpha_{t-1} + \mathbb{D}_0\bar{\alpha} + v_t \qquad v_t \sim \mathbb{N}(0, \Sigma_t) \tag{10.39}$$

where $\alpha_t = vec[A_t(\ell), C_t]$ and $\mathbb{D}_0, \mathbb{D}_1$ are $mk \times mk$ matrices. (10.39) allows for stationary and non-stationary behavior in α_t. For example, the law of motion of the coefficients displays reversion towards the mean $\bar{\alpha}$ if the roots of $\mathbb{D}_1$ are all less than one in absolute value. In principle, Σ_t depends on time, therefore imparting conditional heteroschedastic movements to both the coefficients and the variables of a VAR.

The specification in (10.38)-(10.39) is flexible and can generate a variety of non-linearities in the conditional moment structure. In fact, substituting (10.39) into (10.38) we have

$$y_t = (I_m \otimes \mathsf{X}_t)(\mathbb{D}_1\alpha_{t-1} + \mathbb{D}_0\bar{\alpha}) + (I_m \otimes \mathsf{X}_t)v_t + e_t = X_t\alpha_t^{\dagger} + e_t^{\dagger} \tag{10.40}$$

where $(I_m \otimes \mathsf{X}_t)$ is the matrix of regressors. Depending on the nature of the X_t and the relationship between X_t and v_t, (10.40) encompasses several specifications used in the literature. We consider three such cases in the next example.

Example 10.12 *Suppose $m = 1$, that X_t and v_t are conditionally independent and that var(v_t) $= \Sigma_v$. Then, y_t is conditionally heteroschedastic with mean $X_t\alpha_t^{\dagger}$ and variance $\Sigma_e + X_t'\Sigma_v X_t$. In addition, if X_t includes lagged dependent variables and a constant and $(v_t|X_t) \sim \mathbb{N}(0, \Sigma_v)$, then (10.40) generates a conditionally normal ARMA-ARCH structure. Finally, if X_t includes latent variables or variables which are not perfectly predictable at t, then y_t is non-Gaussian and heteroschedastic (as in Clark's (1973) mixture model).*

Exercise 10.29 *i) Suppose $m = 1$, $X_t = (X_{1t}, X_{2t})$ and assume X_{1t} is correlated with v_t. Show that (10.40) produces a version of the bilinear model of Granger and Anderson (1978). ii) Suppose $v_t = v_{1t} + v_{2t}$, where v_{1t} is independent of X_t and v_{2t} and has covariance matrix Σ_1, and v_{2t} is perfectly correlated with X_t. Show that (10.38)- (10.39) can generate a model with features similar to an ARCH-M model (see Engle, Lilien and Robbins (1987)).*

(10.38)-(10.39) also include, as a special case, Hamilton's (1989) two-state shift model.

Exercise 10.30 *Suppose* $\Delta y_t = a_0 + a_1 \varkappa_t + \Delta y_t^c$ *where* $\varkappa_t = (1-p_2) + (p_1+p_2-1)\varkappa_{t-1} + e_t^x$, e_t^x *is a binomial random variable and* $\Delta y_t^c = A(\ell)\Delta y_{t-1}^c + e_t^c$. *Cast such a model into a TVC framework (Hint: Find its state space format and match coefficients with (10.38)-(10.39)).*

The model can also generate non-normalities in y_t. Typically, such a feature is produced when X_t is a latent variable. However, even when X_t includes only observable variables, e_t and v_t are independently distributed and v_t and X_t conditionally independent, (10.38)-(10.39) can generate non-normalities. To see this set $m = 1$ and define $\hat{e}_{t+\tau} = (\mathbb{D}_1^{\tau+1}\alpha_{t-1} + \mathbb{D}_0\bar{\alpha}\sum_{j=0}^{\tau}\mathbb{D}_1{}^j)'(X_{t+\tau} - E_{t-1}X_{t+\tau}) + (\sum_{j=0}^{\tau-1}\mathbb{D}_1{}^{\tau-j}v_{t+j})'X_{t+\tau} - E_{t-1}(\sum_{j=0}^{\tau-1}\mathbb{D}_1{}^{\tau-j}v_{t+j})'X_{t+\tau}) + v_{t+\tau}'X_{t+\tau} + e_{t+\tau}$.

Exercise 10.31 *Show that, for fixed* t *and all* τ, $E_{t-1}y_{t+\tau} = (\mathbb{D}_1^{\tau+1}\alpha_{t-1} + \mathbb{D}_0\bar{\alpha}\sum_{j=0}^{\tau}\mathbb{D}_1{}^j)'$ $E_{t-1}X_{t+\tau} + E_{t-1}(\sum_{j=0}^{\tau-1}\mathbb{D}_1{}^{\tau-j}v_{t+j})'X_{t+\tau}$; $var_{t-1}y_{t+\tau} = E_{t-1}(\hat{e}_{t+\tau})^2$; $sk_{t-1}(y_{t+\tau}) = \frac{E_{t-1}(\hat{e}_{t+\tau})^3}{(var_{t-1}y_{t+\tau})^{\frac{3}{2}}}$; $kt_{t-1}(y_{t+\tau}) = \frac{E_{t-1}(\hat{e}_{t+\tau})^4}{(var_{t-1}y_{t+\tau})^2}$ *where* sk_{t-1} *and* kt_{t-1} *are the conditional skewness and kurtosis coefficients. Show that for* $\tau = 0$, $sk_{t-1}(y_t) = 0$, $kt_{t-1}(y_t) = 3$, *i.e.* y_t *is conditionally normal.*

For $\tau = 1$ the conditional mean of y_{t+1} is nonlinear and equal to $E_{t-1}(\alpha_{t+1}'X_{t+1}) = (\mathbb{D}_1^2\alpha_{t-1} + \mathbb{D}_0(I+\mathbb{D}_1)\bar{\alpha})'E_{t-1}X_{t+1} + E_{t-1}v_t'\mathbb{D}_1X_{t+1}$ where $E_{t-1}X_{t+1} = [E_{t-1}y_t, y_{t-1}, \ldots y_{t-\ell+1}]$, while its conditional variance is $E_{t-1}((\mathbb{D}_1{}^2\alpha_{t-1} + \mathbb{D}_0\bar{\alpha}(1+\mathbb{D}_1))'(X_{t+1} - E_{t-1}X_{t+1}) + (v_t'\mathbb{D}_1{}'X_{t+1} - E_{t-1}v_t'\mathbb{D}_1{}'X_{t+1}) + v_{t+1}'X_{t+1} + e_{t+1})^2$. Note that $(X_{t+1} - E_{t-1}X_{t+1})' = [e_t^{\dagger}, 0, \ldots, 0]$ and that $((v_t'\mathbb{D}_1{}'X_{t+1}) - E_{t-1}(v_t'\mathbb{D}_1{}'X_{t+1}))$ involves, among other things, terms of the form $v_t'\mathbb{D}_1{}'e_t$. Hence, even when v_t and e_t are normal and independent, y_{t+1} is conditionally non-normal because the prediction errors involve the product of normal random variables. The above argument holds for any $\tau \geq 1$.

10.4.1 Minnesota style prior

If (10.38) is the model for the data and (10.39) the first layer for the prior, we need to specify $\bar{\alpha}$, the evolution of Σ_t and the form of $\mathbb{D}_1$ and $\mathbb{D}_0$. For example, we could use:

$$\mathbb{D}_1 = \phi_0\, I, \qquad \mathbb{D}_0 = I - \mathbb{D}_1 \tag{10.41}$$

$$\bar{\alpha}_{ij\ell} = 1 \qquad \text{if } i = j,\ \ell = 1 \tag{10.42}$$

$$\bar{\alpha}_{ij\ell} = 0 \qquad \text{otherwise} \tag{10.43}$$

$$\Sigma_t = \sigma_t\Sigma_0 \tag{10.44}$$

$$\Sigma_{0ij\ell} = \phi_1\frac{h_1(i,j)}{h_2(\ell)}(\frac{\sigma_j}{\sigma_i})^2 \qquad h_1(i,i) = 1 \tag{10.45}$$

$$\Sigma_{0ij\ell} = \phi_1\phi_4 \qquad \text{if exogenous} \tag{10.46}$$

where $\sigma_t = \phi_3^t + \phi_2\frac{1-\phi_3^{t-1}}{1-\phi_3}$. As in the basic Minnesota prior we assume that Σ_e is fixed, but there is no conceptual difficulty in assuming, e.g., a Wishart prior for Σ_e^{-1}.

With (10.41) the law of motion of the coefficients has a first order autoregressive structure with decay toward the mean. ϕ_0 controls the speed of the decay: for $\phi_0 = 0$ the coefficients are random around $\bar{\alpha}$ and for $\phi_0 = 1$ they are random walks. Higher order processes can be obtained by substituting the identity matrix in (10.41) with an appropriate matrix. The prior mean and the prior variance for the time zero coefficients are identical to those of the basic Minnesota prior except that we allow a general pattern of weights for different variables in different equations via the function $h_1(i,j)$. The variance of the innovation in the coefficients evolves linearly. The nature of time variations can be clearly understood using: $\Sigma_t = V_0\Sigma_0 + V_1\Sigma_{t-1}$, which has the same structure as the law of motion of the coefficients, and which reduces to the expression in (10.44) if $V_0 = \phi_2 \times I, V_1 = \phi_3 \times I$. For $\phi_3 = 0$ the coefficients are time varying but no heteroschedasticity is allowed, while for $\phi_2 = 0$ the variance of the coefficients is geometrically related to Σ_0. Finally, if $\phi_2 = \phi_3 = 0$, time variations and heteroschedasticity are absent.

Empirical Bayes methods can be employed to estimate the hyperparameters ϕ on a training sample of data going from $(-\tau, 0)$. As usual, the predictive density can be constructed and evaluated numerically using the Kalman filter.

Exercise 10.32 *Write down the predictive density for the TVC-VAR model. Specify exactly how to use the Kalman filter to numerically maximize the predictive density.*

Posterior inference can be conducted conditional on the estimates of ϕ, i.e., we use $g(\alpha|y, \hat{\phi}_{ML-II}) \propto f(y|\alpha)g(\alpha|\hat{\phi}_{ML-II})$ in place of $g(\alpha|y)$. Note that while the full posterior averages over all possible values of ϕ, the empirical-Bayes posterior uses ML-II estimates. Clearly, if $f(y|\phi)$ is flat in the hyperparameter space, differences will be minor.

Example 10.13 *Continuing with example 10.5, we add time variations to the coefficients of the BVAR and forecast inflation using the same style of Minnesota prior outlined above, but set $\phi_3 = 0$. We use a simplex algorithm to maximize the predictive density with respect to ϕ's. The optimal values are $\phi_0 = 0.98$, $\phi_1 = 0.11, \phi_2 = 0.1e-8, \phi_4 = 1000$, while $h_1(i,j) = 0.4\ \forall i,j,\ h_2(\ell) = \ell^{0.4}$. The Theil-U statistics one year ahead are 0.93 for the sample 1996:1-2000:4 and 0.89 for the sample 1991:1-1995:4 (the posterior standard error is 0.03 in both cases). Therefore, time variations in the coefficients appear to be important in forecasting Italian inflation. However, time variations in the variance hardly matter. In fact, setting $\phi_2 = 0$, the Theil-U are 0.95 and 0.90, respectively.*

Exercise 10.33 *(Ciccarelli and Rebucci) Suppose $y_{1t} = A_{11}(\ell)y_{1t-1} + y_{2t}A_{12}$ and $y_{2t} = A_{22}(\ell)y_{1t-1} + v_t$ and suppose a researcher estimates $y_{1t} = A(\ell)y_{1t-1} + e_t$.*
i) Show that $A_{ols}(\ell)$ is biased unless $A_{22}(\ell) = 0$.
ii) Consider the approximating model $y_{1t} = A(\ell)y_{1t-1} + A^c(\ell)y_{1t-1} + e_t$ where $A^c(\ell) = A_{22}(\ell)A_{12}$ and $e_t = v_tA_{12}$. Clearly, the estimated model sets $A^c(\ell) = 0$, otherwise perfect collinearity would result. Suppose $\alpha = vec(A^c(\ell), A(\ell)) \sim \mathbb{N}(\bar{\alpha}, \bar{\Sigma}_\alpha)$ where $\bar{\alpha} = (0, \bar{\alpha}_2)$ and $\bar{\Sigma}_\alpha = diag[\bar{\Sigma}_{\alpha_1}, \bar{\Sigma}_{\alpha_2}]$. Show that $g(\alpha|y) \sim \mathbb{N}(\tilde{\alpha}, \tilde{\Sigma}_\alpha)$. Show the form of $\tilde{\alpha}, \tilde{\Sigma}_\alpha$. In particular, show that, in the formula for the posterior mean, the OLS estimator receives less weight

than in standard problems. Show that the posterior for $A^c(\ell)$ is centered away from zero to correct for the skewness produced by omitting a set of regressors. How would your answer change if coefficients are functions of time?

10.4.2 Hierarchical prior

A BVAR with time varying coefficients is a state space model where the coefficients (variances) play the role of the unobservable states. Full hierarchical estimation of such models do not present difficulties once it is understood that time-varying and time invariant features can be jointly estimated. The Gibbs sampling is particularly useful for this purpose.

Here we consider a simple version of the model (10.38)-(10.39) and leave the discussion of a more complicated setup to a later section. The specification we employ has the form:

$$\begin{aligned} y_t &= X_t\alpha_t + e_t \qquad & e_t &\sim \mathbb{N}(0,\Sigma_e) \\ \alpha_t &= \mathbb{D}_1\alpha_{t-1} + v_t \qquad & v_t &\sim \mathbb{N}(0,\Sigma_a) \end{aligned} \tag{10.47}$$

where $X_t = (I_m \otimes \mathsf{X}_t)$. We assume that $\mathbb{D}_1$ is known and discuss in an exercise how to estimate it, in the case it is not. Posterior draws from the distribution of the unknown parameters (Σ_e, Σ_a) and of the unobserved state $\{\alpha_t\}_{t=1}^T$ can be obtained with the Gibbs sampler. Let $\alpha^t = (\alpha_0, \ldots, \alpha_t)$, $y^t = (y_0, \ldots, y_t)$. To use the Gibbs sampler we need three conditional posteriors: $(\Sigma_a|y^t, \alpha^t, \Sigma_e)$, $(\Sigma_e|y^t, \alpha^t, \Sigma_a)$ and $(\alpha^t|y^t, \Sigma_e, \Sigma_a)$.

Suppose that $g(\Sigma_e^{-1}, \Sigma_a^{-1}) = g(\Sigma_e^{-1})g(\Sigma_a^{-1})$ and that each is Wishart with $\bar\nu_0$ and $\bar\nu_1$ degrees of freedom and scale matrices $\bar\Sigma_e, \bar\Sigma_a$, respectively. Then, since e_t, v_t are normal

$$(\Sigma_e^{-1}|y^t, \alpha^t, \Sigma_a^{-1}) \sim \mathbb{W}(\bar\nu_0 + T, (\bar\Sigma_e^{-1} + \sum_t (y_t - X_t\alpha_t)(y_t - X_t\alpha_t)')^{-1})$$

$$(\Sigma_a^{-1}|y^t, \alpha^t, \Sigma_e^{-1}) \sim \mathbb{W}(\bar\nu_1 + T, (\bar\Sigma_a^{-1} + \sum_t (\alpha_t - \mathbb{D}_1\alpha_{t-1})(\alpha_t - \mathbb{D}_1\alpha_{t-1})')^{-1})$$

To obtain the conditional posterior of α^t notice that $g(\alpha^t|y^t, \Sigma_e, \Sigma_a) = g(\alpha_t|y^t, \Sigma_e, \Sigma_a)$ $g(\alpha_{t-1}|y^t, \alpha_t, \Sigma_e, \Sigma_a) \cdots g(\alpha_0|y^t, \alpha_1, \Sigma, V)$. Therefore, a sequence α^t can be obtained drawing each element from the corresponding conditional posterior while α_t is drawn from the marginal $g(\alpha_t|y^t, \Sigma_e, \Sigma_a)$. Let $\alpha_\tau^t = (\alpha_\tau, \ldots, \alpha_t)$ and $y_\tau^t = (y_\tau, \ldots, y_t)$. Then

$$\begin{aligned} g(\alpha_\tau|y^t, \alpha_{\tau+1}^t, \Sigma_e, \Sigma_a) &\propto g(\alpha_\tau|y^\tau, \Sigma_e, \Sigma_a) g(\alpha_{\tau+1}|y^\tau, \alpha_\tau, \Sigma_e, \Sigma_a) \\ &\times f(y_{\tau+1}^t, \alpha_{\tau+1}^t|y_\tau, \alpha_\tau, \alpha_{\tau+1}, \Sigma_e, \Sigma_a) \\ &= g(\alpha_\tau|y^\tau, \Sigma_e, \Sigma_a) g(\alpha_{\tau+1}|\alpha_\tau, \Sigma_e, \Sigma_a) \end{aligned} \tag{10.48}$$

The first two terms involve posterior distributions obtained with data up to τ and the last term the distribution of the data and the coefficients from $\tau+1$ until t. The last line follows from the fact that α_τ is independent of $y_{\tau+1}^t, \alpha_{\tau+1}^t$, conditional on $(y^\tau, \Sigma_e, \Sigma_a)$. It is immediate to recognize that the two densities in (10.48) can be computed from the smoothing and the predictive equations of the Kalman filter (see chapter 6). Let $\alpha_{t|t} \equiv E(\alpha_t|y^t, \Sigma_e, \Sigma_a) = \alpha_{t|t-1} + \mathsf{K}_t(y_t - X_t\alpha_{t|t-1})$; $\Sigma_{t|t} \equiv var(\alpha_t|y^t, \Sigma_a, \Sigma_e) = (I - \mathsf{K}_t X_t)\Sigma_{t|t-1}$ where $\alpha_{t|t-1} =$

$\mathbb{D}_1\alpha_{t-1|t-1}$, $\mathsf{K}_t = \Sigma_{t|t-1}X_t'(X_t\Sigma_{t|t-1}X_t' + \Sigma_e)^{-1}$, and $\Sigma_{t|t-1} \equiv var(\alpha_t|y^{t-1}, \Sigma_e, \Sigma_a) = \mathbb{D}_1\Sigma_{t-1|t-1}\mathbb{D}_1' + \Sigma_a$. Using the linearity of the model and the Gaussian structure of (10.47), $g(\alpha_\tau|y^\tau, \Sigma_e, \Sigma_a)$ is normal with mean $\alpha_{\tau|\tau}$ and variance $\Sigma_{\tau|\tau}$, while $g(\alpha_{\tau+1}|y^\tau, \alpha_\tau, \Sigma_e, \Sigma_a)$ is normal with mean $\mathbb{D}_1\alpha_\tau$ and variance Σ_a. Therefore, given a prior for α_0, all conditional densities are Gaussian and to keep track of these distributions we only need to update conditional means and variances. Hence, to draw samples from $g(\alpha^t|y^t, \Sigma, \Sigma_a)$ we use the following:

Algorithm 10.3

1) *Run the Kalman filter, save* $\alpha_{t|t}$, $\Sigma_t = \Sigma_{t|t} - \mathbb{M}_t\Sigma_{t+1|t}\mathbb{M}_t'$, *and* $\mathbb{M}_t = \Sigma_{t|t}\Sigma_{t+1|t}^{-1}$.

2) *Draw* $\alpha_t^l \sim \mathbb{N}(\alpha_{t|t}, \Sigma_{t|t})$, $\alpha_{t-j}^l \sim \mathbb{N}(\alpha_{t-j|t-j} + \mathbb{M}_{t-j}(\alpha_{t-j+1}^l - \alpha_{t-j|t-j}), \Sigma_{t-j})$, $j \geq 1$.

3) *Repeat* $l = 1, \ldots L$ *times*

It is straightforward to allow for an unknown $\mathbb{D}_1$ and a time-varying Σ_a.

Exercise 10.34 *Assume that* $\mathbb{D}_1$ *is unknown and assume a normal prior on its nonzero elements i.e.* $\mathbb{D}_1^0 \sim \mathbb{N}(\bar{\mathbb{D}}_1, \bar{\sigma}_{D_1}^2)$. *Show that* $g(\mathbb{D}_1^0|\alpha^t, y^t, \Sigma_e, \Sigma_a) \sim \mathbb{N}((\alpha_{t-1}'\Sigma_a^{-1}\alpha_{t-1} + \sigma_{D_1}^{-2})^{-1}$ $(\alpha_{t-1}'\Sigma_a^{-1}\alpha_t + \sigma_{D_1}^{-2}\bar{\mathbb{D}}_1); (\alpha_{t-1}'\Sigma_a^{-1}\alpha_{t-1} + \sigma_{D_1}^{-2})^{-1})$.

Exercise 10.35 *Let* $\Sigma_{at} = \sigma_t\Sigma_a$. *How would you construct the conditional posterior distribution for* Σ_{at}? *(Hint: treat* σ_t *as a parameter and assume a conjugate prior).*

The next extension is useful to compute the likelihood of DSGE models which are not linearized around the steady state.

Exercise 10.36 *(Non-linear state space models) Consider the state space model:*

$$\begin{aligned} y_t &= f_{1t}(\alpha_t) + e_t \qquad & e_t \sim \mathbb{N}(0, \Sigma_e) \\ \alpha_t &= f_{2t}(\alpha_{t-1}) + v_t \qquad & v_t \sim \mathbb{N}(0, \Sigma_a) \end{aligned} \tag{10.49}$$

where f_{1t} *and* f_{2t} *are given but perhaps depend on unknown parameters. Show that* $(\alpha_t|\alpha_{j\neq t}, \Sigma_e, \Sigma_a, y^t) \propto \mathsf{h}_1(\alpha_t)\mathsf{h}_2(\alpha_t)\mathbb{N}(f_{2t}(\alpha_{t-1}), \Sigma_a)$ *where* $\mathsf{h}_1(\alpha_t) = \exp\{-0.5(\alpha_{t+1} - f_{2t}(\alpha_t))'$ $\Sigma_a^{-1}(\alpha_{t+1} - f_{2t}(\alpha_t))\}$; $\mathsf{h}_2(\alpha_t) = \exp\{-0.5(y_t - f_{1t}(\alpha_t))'\Sigma_e^{-1}(y_t - f_{1t}(\alpha_t))\}$. *Describe how to use an acceptance sampling algorithm to draw from this posterior distribution.*

Finally, we consider the case of non-normal errors. While for macroeconomic data the assumption of normality is, by and large, appropriate, for robustness purposes it may be useful to allow for non-normalities. As noted, the conditional moments of (10.47) are nonlinear for $\tau \geq 1$. To generate non-normalities, when $\tau = 0$, it is sufficient to add a nuisance parameter ϕ_5 to the variance of the error term, i.e., $(\alpha_t|\alpha_{t-1}, \phi_5, \Sigma_a) \sim \mathbb{N}(\mathbb{D}_1\alpha_{t-1}, \phi_5\Sigma_a)$ where $g(\phi_5)$ is chosen to mimic a distribution of interest. For example, suppose that ϕ_5 is exponentially distributed with mean equal to 2. Since $g(\alpha_t|\alpha_{t-1}, \Sigma_a, \phi_5)$ is normal with mean $\mathbb{D}_1\alpha_{t-1}$ and

variance $\phi_5\Sigma_a$; $g(\phi_5|y^t,\alpha^t,\Sigma_a) \propto \sqrt{\frac{1}{\phi_5}}\exp\{-0.5[\phi_5+(\alpha_t-\mathbb{D}_1\alpha_{t-1})'\phi_5^{-1}\Sigma_a^{-1}(\alpha_t-\mathbb{D}_1\alpha_{t-1})]\}$ which is the kernel of the generalized inverse Gaussian distribution. A similar approach can be used to model non-normalities in the measurement equation.

Exercise 10.37 *Suppose* $(y_t|\alpha_t,x_t,\phi_6,\Sigma_e) \sim \mathbb{N}(x_t\alpha_t,\phi_6\Sigma_e)$ *and that* $g(\phi_6)$ *is* $exp(2)$. *Show the form of the conditional posterior for* ϕ_6. *Describe how to draw sequences for* ϕ_6.

Exercise 10.38 *Let* $y_t = x_t\alpha_t,\ t=1,\dots T$ *where conditional on* x_t $\alpha_t' = (\alpha_{1t},\dots\alpha_{kt})$*is iid with mean* $\bar{\alpha}$ *and variance* $\bar{\Sigma}_\alpha$, $|\bar{\Sigma}_\alpha| \neq 0$. *Assume that* $\bar{\alpha}$ *and* $\bar{\Sigma}_\alpha$ *are known and let* $\alpha = (\alpha_1,\dots,\alpha_t)$.
i) Show that the minimum MSE estimator of α *is* $\tilde{\alpha} = (I_T\otimes\bar{\Sigma}_\alpha)x'\Omega^{-1}y + (I_{Tk} - (I_T\otimes\bar{\Sigma}_\alpha)x'\Omega^{-1}x)(\mathbf{1}\otimes\bar{\alpha})$ *where* $\Omega = x(I_T\otimes\Sigma_\alpha)x'$,$x = diag(x_1',\dots x_t')$ *and* $\mathbf{1} = [1,\dots,1]'$.
ii) Show that if $\bar{\alpha} = \alpha_0 + v_a$, $v_a \sim (0,\Sigma_{\bar{a}})$ *and* $\Sigma_{\bar{a}}$ *is known, the best minimum MSE estimator of* $\bar{\alpha}$ *equals* $(x'\Omega^{-1}x + \Sigma_{\bar{a}}^{-1})^{-1}(x'\Omega^{-1}y + \Sigma_{\bar{a}}^{-1}\alpha_0)$. *Show that as* $\Sigma_{\bar{a}} \to \infty$ *the optimal MSE estimator is the GLS estimator.*

Exercise 10.39 *(Cooley and Prescott) Let* $y_t = x_t\alpha_t$ *where* x_t *is a* $1\times k$ *vector;* $\alpha_t = \alpha_t^P + e_t$; $\alpha_t^P = \alpha_{t-1}^P + v_t$ *where* $e_t \sim (0,(1-\varrho)\sigma^2\Sigma_e)$, $v_t \sim (0,\varrho\sigma^2\Sigma_v)$ *and assume* Σ_e, Σ_v *known. Here* ϱ *represents the speed of adjustment of* α_t *to structural changes (for* $\varrho \to 1$ *permanent changes are large relative to transitory ones). Let* $y = [y_1,\dots,y_T]'$, $x = [x_1,\dots,x_T]'$ *and* $\alpha^p = (\alpha_{1t}^p,\dots,\alpha_{kt}^p)'$.
i) Show that the model is equivalent to $y_t = x_t'\alpha_t^P + \epsilon_t$; $\epsilon_t \sim (0,\sigma^2\Omega(\varrho))$. *Display* $\Omega(\varrho)$.
ii) Show that, conditional on ϱ, *the minimum MSE estimators for* (α^p,σ^2) *are* $\alpha_{ML}^p(\varrho) = (x'\Omega(\varrho)^{-1}x)^{-1}\ (x'\Omega(\varrho)^{-1}y)$ *and* $\sigma_{ML}^2(\varrho) = \frac{1}{T}(y-x\alpha_{ML}^p(\varrho))'\Omega(\varrho)^{-1}(y-x\alpha_{ML}^p(\varrho))$. *Describe a way to maximize the concentrated likelihood as a function of* ϱ.
iii) Obtain posterior estimators for $(\alpha,\varrho,\sigma^2)$ *when* $g(\alpha,\varrho,\sigma^2)$ *is non-informative. Set up a Gibbs sampler algorithm to compute the joint posterior of the three parameters.*

10.5 Panel VAR models

We have extensively discussed macro panel data in chapter 8. Therefore, the focus of this section is narrow. Our attention centers on three problems. First, how to specify Bayesian univariate dynamic panels. Second, how to dynamically group units in the cross section. Third, how to setup panel VAR models with cross sectional interdependencies. Univariate dynamic panels emerge, for example, when estimating steady state income per-capita, or when examining the short and long run effects of oil shocks on output across countries. Grouping is particularly useful, for example, if one is interested in knowing if there are countries which react differently than others after e.g. financial crises. Finally, models with interdependencies are useful to study a variety of transmission issues across countries or sectors which can not be dealt with the models of chapter 8.

10.5.1 Univariate dynamic panels

For $i = 1, \dots n$, the model we consider is:

$$y_{it} = A_{1i}(\ell) y_{it-1} + \bar{y}_i + A_{2i}(\ell) Y_t + e_{it} \quad e_{it} \sim (0, \sigma_i^2) \tag{10.50}$$

where $A_{ji}(\ell) = A_{ji1}\ell + \ldots + A_{jiq_j}\ell^{q_j}$ $j = 1, 2$ and $\bar{y}_i$ is the unit specific fixed effect. Here Y_t includes variables which account for cross sectional interdependencies. For example, if y_{it} are regional sales, one element of Y_t could be a national business cycle indicator. Because variables like Y_t are included, $E(e_{it}e_{j\tau}) = 0 \; \forall i \neq j$, all t, τ. We can calculate a number of statistics from (10.50). For example, long run multipliers to shocks are $(1 - A_{1i}(1))^{-1}$ and long run multipliers to changes in Y_t are $(1 - A_{1i}(1))^{-1} A_{2i}(1)$.

Example 10.14 *Let y_{it} be output in Latin American country i and let $Y_t = (x_{1t}, i_t)$, where i_t is US interest rate. Suppose $i_t = A_3(\ell)\epsilon_t$. Then $(1 - A_{1i}(\ell)^{-1} A_{2i}(\ell) A_3(\ell)$ traces out the effect of unitary US interest rate shock at t on the output of country i from t on.*

Stacking the T observations for (y_{it}, Y_t, e_{it}) and the fixed effect into the vectors $(y_i, Y, e_i, \mathbf{1})$, letting $\mathsf{X}_i = (y_i, Y, \mathbf{1})$, $\Sigma_i = \sigma_i^2 \times I_T$, $\alpha = [A_1, \ldots A_n]'$, $A_i = (A_{1i1}, \ldots, A_{iq_1}, \bar{y}_i, A_{1i1}, \ldots, A_{2iq_2})$ and setting $y = (y_1, \ldots y_n)', e = (e_1, \ldots, e_n)'$:

$$y = (I_n \otimes \mathsf{X}_i)\alpha + e \quad e \sim (0, \Sigma_i \otimes I_n) \tag{10.51}$$

Clearly, (10.51) has the same format as a VAR, except that X_i are unit specific and the covariance matrix of the shocks has a diagonal heteroschedastic structure. The first feature is due to the fact that we do not allow for interdependencies across units. The latter is easy to deal with once (10.51) is transformed so that the innovations have spherical disturbances.

If e is normal, the likelihood function of a univariate dynamic panel is therefore the product of a normal for α, conditional on $\Sigma_i \otimes I_n$, and n Gamma densities for Σ_i^{-1}. Since the variance of e is diagonal, α_{ML} can be obtained equation by equation.

Exercise 10.40 *Show that α_{ML} obtained from (10.51) is the same as the estimator obtained by stacking weighted least square estimators obtained from (10.50) for each i.*

Conjugate priors for dynamic panels are similar to those described in section 10.2. Since $var(e)$ is diagonal, we can choose $\sigma_i^{-2} \sim \mathbb{G}(a_1, a_2)$, each i. Given the panel framework we can use the exchangeability assumption if, a-priori, we expect the A_i to be similar across units. An exchangeable prior on A_i takes the form $A_i \sim \mathbb{N}(\bar{A}, \bar{\sigma}_A^2)$ where $\bar{\sigma}_A^2$ measures the degree of heterogeneity an investigator expects to find in the cross section.

Exercise 10.41 *(Lindlay and Smith) Suppose the model (10.50) has k coefficients in each equation and that $A_i = \bar{A} + v_i$, $i = 1, \ldots n$, $v_i \sim \mathbb{N}(0, \bar{\sigma}_A^2)$, where $\bar{A}, \bar{\sigma}_A^2$ are known. Show the form of the posterior mean for A_i. Assuming that σ_i^2 is fixed, show the form of the posterior variance for A_i. Argue that the posterior mean for the stacked vector of A_i is the same as the one obtained by calculating the posterior mean for the system (10.51).*

Exercise 10.41 highlights the importance of exchangeable priors in a model like (10.51). In fact, exchangeability preserves independence across equations and the posterior mean of the coefficients of a dynamic panel can be computed equation by equation.

Exercise 10.42 *(Canova and Marcet) Suppose you want to set up an exchangeable prior on the difference of the coefficients across equations, i.e.* $\alpha_i - \alpha_j \sim \mathbb{N}(0, \Sigma_a)$. *This is advantageous since there is no need to specify the prior mean* $\bar{\alpha}$. *Show the structure of* Σ_a *which insures that the ordering of the units in the cross section does not matter.*

We already mentioned the pooling dilemma in section 4 of Chapter 8. We return to this problem in the next exercise which gives conditions under which the posterior distribution for A_i reflects prior, pooled and/or single unit sample information.

Exercise 10.43 *(Zellner and Hong) Let* $y_i = x_i\alpha_i + e_i, \;\; i = 1, \dots n$ *where* x_i *may include lags of* y_{it} *and for each* i, y_i *is a* $T \times 1$ *vector,* x_i *a* $T \times k$ *vector and* α_i *a* $k \times 1$ *vector and* $e_i \sim$ iid $\mathbb{N}(0, \sigma_e^2)$. *Assume that* $\alpha_i = \bar{\alpha} + v_i$, *where* $v_i \sim$ iid $\mathbb{N}(0, \kappa^{-1}\sigma_v^2 I_k)$ *with* $0 < \kappa \leq \infty$.
(i) Show that a conditional point estimate for $\alpha = (\alpha_1', \dots, \alpha_N')'$ *is the* $Nk \times 1$ *vector* $\tilde{\alpha} = (x'x + \kappa I_{nk})^{-1}(x'x\alpha_{ols} + \kappa \mathbb{I}\alpha_p)$ *where* $x = \text{blockdiag}\{x_i\}$; $\alpha_{ols} = (x'x)^{-1}(x'y)$; $y = (y_1', \dots, y_N')'$, $\alpha_{ols} = (\alpha_{1,ols}', \dots \alpha_{N,ols}')'$, $\alpha_{i,ols} = (x_i'x_i)^{-1}(x_i'y_i)$, $\mathbb{I} = (I_k, \dots I_k)$, $\alpha_p = (\sum_i x_i'x_i)^{-1} (\sum_i x_i'x_i\alpha_{i,ols})$. *Conclude that* $\tilde{\alpha}$ *is a weighted average of individual OLS estimates and of the pooled estimate* α_p. *Show that, as* $\kappa \to \infty$, $\tilde{\alpha} = \alpha_p$.
(ii) (g-prior) Assume that $v_i \sim$ iid $\mathbb{N}(0, (x_i'x_i)^{-1}\sigma_v^2)$. *Show that* $\tilde{\alpha}_i^1 = (\alpha_{i,ols} + \frac{\sigma_e^2}{\sigma_v^2}\bar{\alpha})/(1 + \frac{\sigma_e^2}{\sigma_v^2})$ *Conclude that* $\tilde{\alpha}_i^1$ *is a weighted average of the OLS estimate and the prior mean* $\bar{\alpha}$.
(iii) Show that if $g(\bar{\alpha})$ *is non-informative,* $\tilde{\alpha}_i^2 = (\alpha_{i,ols} + \frac{\sigma_e^2}{\sigma_v^2}\alpha_p)/(1 + \frac{\sigma_e^2}{\sigma_v^2})$. *Conclude that, as* $\frac{\sigma_e^2}{\sigma_v^2} \to \infty$, $\tilde{\alpha}_i = \alpha_p$ *and, as* $\frac{\sigma_e^2}{\sigma_v^2} \to 0$, $\tilde{\alpha}_i = \alpha_{i,ols}$.

Next we describe how dynamic univariate panels can be used to estimate the steady state distribution of income per-capita and of the convergence rates in a panel of EU regions.

Example 10.15 *Here* $A_{1i}(\ell)$ *has only one non-zero element (the first one),* Y_t *is the average EU GDP per-capita and* $A_{2ij} = 1$ *if* $j = 0$ *and zero otherwise. Hence (10.50) is:*

$$\ln(\frac{y_{it}}{Y_t}) = \bar{y}_i + A_i \ln(\frac{y_{it-1}}{Y_{t-1}}) + e_{it} \qquad e_{it} \sim \mathbb{N}(0, \sigma_i^2) \tag{10.52}$$

We let $\alpha_i = (\bar{y}_i, A_i)$ *and assume* $\alpha_i = \bar{\alpha} + v_i$, *where* $v_i \sim \mathbb{N}(0, \sigma_a^2 I)$.

We treat σ_i^2 *as known (and estimate it from invidual OLS regressions), assume* $\bar{\alpha}$ *known (estimated averaging individual OLS estimates) and treat* $\bar{\sigma}_a^2$ *as fixed. Let* $\frac{\sigma_i^2}{\sigma_{a_j}^2}$, $j = 1, 2$ *measures the relative importance of prior and sample information: if this ratio goes to infinity sample information does not matter; viceversa, if it is close to zero, prior information is irrelevant. We choose a relative loose prior (*$\frac{\sigma_i^2}{\sigma_{a_j}^2} = 0.5, j = 1, 2$*). Using income per-capita*

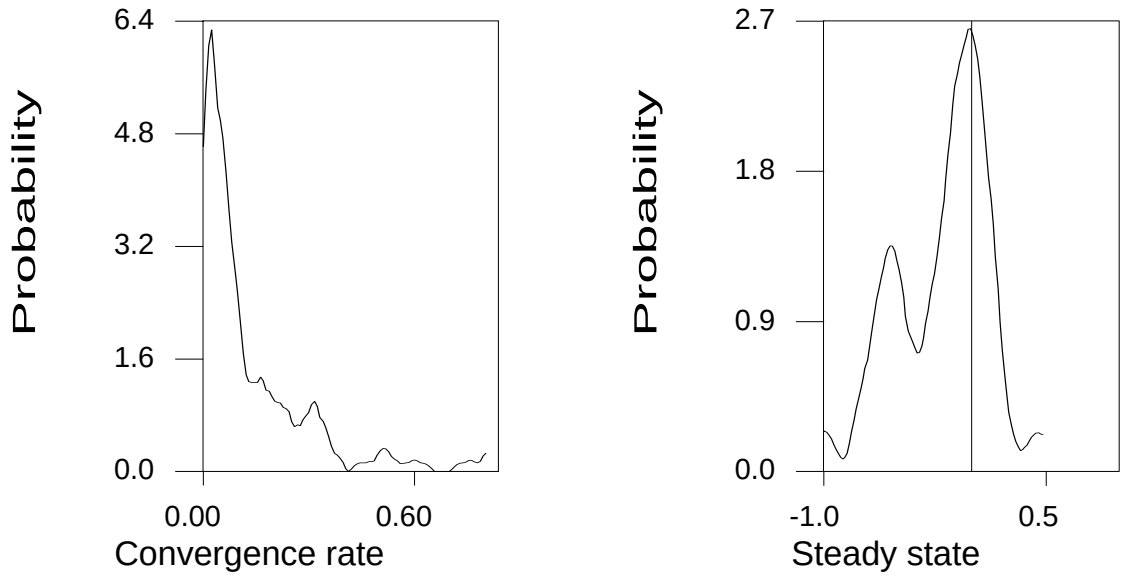


Figure 10.4: Cross sectional distributions.

for 144 EU regions from 1980 to 1996 we calculate the relative steady state for unit i using $\tilde{SS}_i = \tilde{\bar{y}}_i \frac{1-\tilde{A}_i^T}{1-\tilde{A}_i} + \tilde{A}_i^{T+1} \frac{y_{i0}}{Y_0}$ *where* $\tilde{\bar{y}}_i, \tilde{A}_i$ *are posterior mean estimates. The rate of convergence to the steady state is* $\tilde{CV}_i = 1 - \tilde{A}_i$ *(If* $\tilde{A} > 1$*, we set* $\tilde{CV} = 0$*). We plot the cross sectional distribution of* $\tilde{CV}$ *and* $\tilde{SS}$ *in figure 10.4. The mode of the convergence rate is 0.09, implying much faster catch up than the literature has found (see e.g. Barro and Sala (1995)). The highest 95% credible set is however large (it goes from 0.03 to 0.55). The cross sectional distribution of relative steady states has at least two modes: one at low relative levels of income and one just below the EU average.*

At times, when the panel is short, one wishes to use cross-sectional information to get better estimates of the parameters of each unit. In other cases, one is interested in estimating the average cross sectional effect. In both situations, the tools of Meta analysis come handy.

Example 10.16 *Continuing with example 10.15, suppose* $g(SS_i) \sim \mathbb{N}(\bar{SS}, \sigma_{SS}^2)$ *where* $\sigma_{SS} = 0.4$ *and assume* $g(\bar{SS}) \propto 1$*. Using the logic of hierarchical models,* $g(\bar{SS}|y)$ *combines prior and data information and* $g(SS_i|y)$ *combines unit specific and pooled information. The posterior mean for* $\bar{SS}$ *is -0.14 indicating that the distribution is highly skewed to the left, the variance is 0.083 and a credible 95 percent interval is (-0.30, 0.02). Since a credible 95 percent posterior interval for* SS_i *is (-0.51, 0.19), this posterior distribution largely overlaps with the one in figure 10.4.*

10.5.2 Endogenous grouping

There are many situations when one would like to know whether there are groups in the cross section of a dynamic panel. For example, one type of growth theory predicts the existence of convergence clubs, where clubs are defined by similarities in the features of the various economies or government policies. In monetary economics, one is typically interested in knowing whether regional economies respond differently to union wide monetary policy disturbances or whether the behavioral responses of certain groups of agents (credit constrained vs. credit unconstrained consumers, large vs. small firms, etc.) can be identified. In general, these classifications are exogenously chosen (see for example, Gertler and Gilchrist (1991)) and somewhat arbitrary.

In this subsection we describe a procedure which simultaneously allows for endogenous grouping of cross sectional units and for Bayesian estimation of the parameters of the model. The basic idea is simple: if units i and i' belong to a group, the vector of coefficients will have the same mean and the same dispersion but if the don't, the vector of coefficients of the two units will have different moments.

Let n be the size of the cross section, T the size of the time series, and $\mathcal{O} = 1, 2, \ldots n!$ the ordering of the units of the cross section (the ordering producing a group is unknown). We assume there could be $\psi = 1, 2, \ldots, \bar{\psi}$ break points, $\bar{\psi}$ given. For each group $j = 1, \ldots, \psi+1$ and each unit $i = 1, \ldots, n^j(\mathcal{O})$

$$y_{it} = \bar{y}_i + A_{1i}(\ell) y_{it-1} + A_{2i}(\ell) Y_{t-1} + e_{it} \quad e_{it} \sim (0, \sigma^2_{e_i}) \tag{10.53}$$

$$\alpha_i^j = \bar{\alpha}^j + v_i^j \qquad v_i^j \sim (0, \bar{\Sigma}_j) \tag{10.54}$$

where $\alpha_i = [\bar{y}_i, A_{1i1}, \ldots, A_{1iq_1}, A_{2i1}, \ldots, , A_{2iq_2}]'$ is the $k_i \times 1$ vector of coefficients of unit $i, k_i = q_1 + q_2 + 1$, $n^j(\mathcal{O})$ is the number of units in group j, given the $\mathcal{O}$-th ordering, $\sum_j n^j(\mathcal{O}) = n$, for each $\mathcal{O}$. In (10.54), α_i is random but the coefficients of the $n^j(\mathcal{O})$ units belonging to group j have the same mean and same covariance matrix. Since the exchangeable structure may differ across groups, (10.53)-(10.54) capture the idea that there may be clustering of units within groups but that groups may drift apart.

The alternative to (10.53)-(10.54) is a model with homogeneous dynamics in the cross section, that is $\bar{\psi} = 0$, and an exchangeable structure for all units of the cross section, i.e.

$$\alpha_i = \bar{\alpha} + v_i \qquad i = 1, \ldots, n \quad v_i \sim (0, \bar{\Sigma}_i) \tag{10.55}$$

Let Y be a $(nTm) \times 1$ the vector of left hand side variables in (10.53) ordered to have the n cross sections for each $t = 1, \ldots T$, m times, X be a $(nTm) \times (nk)$ matrix of the regressors, α be a $(nk) \times 1$ vector of coefficients, E a $(nTm) \times 1$ vector of disturbances, $\bar{\alpha}$ a $(\psi+1)k \times 1$ vector of means of α, A be a $(nk) \times (\psi+1)k$ matrix, $A = diag\{A_j\}$, where A_j has the form $\mathbf{1} \otimes I_k$ where I_k is a $k \times k$ identity matrix and $\mathbf{1}$ is a $n^j(\mathcal{O}) \times 1$ vector of ones. Given an ordering $\mathcal{O}$, the number of groups ψ, and the location of the break point $h^j(\mathcal{O})$, we can rewrite (10.53) – (10.54) as:

$$Y = X\alpha + E \qquad E \sim (0, \Sigma_E) \tag{10.56}$$

$$\alpha = \Xi\bar{\alpha} + V \qquad V \sim (0, \Sigma_V) \tag{10.57}$$

where Σ_E is $(nTm) \times (nTm)$ and $\Sigma_V = diag\{\Sigma_i\}$ is a $(nk) \times (nk)$ matrix and Ξ is a matrix of zeros and ones. To complete the specification we need priors for $(\bar{\alpha}, \Sigma_E, \Sigma_V)$ and for the submodel characteristics $\mathcal{M}$, indexed by $(\mathcal{O},\ \psi,\ h^j(\mathcal{O}))$. Since the calculation of the posterior distribution is complicated, we take an Empirical Bayes approach.

The approach to group units proceeds in three steps. Given $(\bar{\alpha},\ \Sigma_E,\ \Sigma_V, \mathcal{O})$, we examine how many groups are present. Given $\mathcal{O}$ and $\hat{\psi}$, we check for the location of the break points. Finally we iterate on the first two steps, altering $\mathcal{O}$. The selected submodel is the one that maximizes the predictive density over orderings $\mathcal{O}$, groups ψ, and break points $h^j(\mathcal{O})$.

Let $f(Y|H_0)$ be the predictive density of the data under cross sectional homogeneity. Furthermore, let I^ψ be the set of possible break points when there are ψ groups. Let $f(Y^j|H_\psi, h^j(\mathcal{O}), \mathcal{O})$ be the predictive density for group j, under the assumption that there are ψ break points with location $h^j(\mathcal{O})$, using ordering $\mathcal{O}$ and let $f(Y|H_\psi, h^j(\mathcal{O}), \mathcal{O}) = \prod_{j=1}^{\psi+1} f(Y^j|H_\psi, h^j(\mathcal{O}), \mathcal{O})$. Define the quantities

- $f^-(Y|H_\psi, \mathcal{O}) \equiv \sup_{h^j(\mathcal{O}) \in I^\psi} f(Y|H_\psi, h^j(\mathcal{O}), \mathcal{O})$,
- $f^\dagger(Y|H_\psi) \equiv \sup_{\mathcal{O}} f^-(Y|H_\psi, \mathcal{O})$,
- $f^0(Y|H_\psi, \mathcal{O}) \equiv \sum_{h^j(\mathcal{O}) \in I^\psi} g_i^j(\mathcal{O}) f(Y|H_\psi, h^j(\mathcal{O}), \mathcal{O})$,

where $g_i^j(\mathcal{O})$ is the prior probability that there is a break at location $h^j(\mathcal{O})$ for group j of ordering $\mathcal{O}$. f^- gives the maximized predictive density with respect to the location of break points, for each ψ and $\mathcal{O}$; $f^\dagger$ the maximized predictive density, for each ψ, once the location of the break point and the ordering of the data are chosen optimally. f^0 gives the average predictive density with ψ breaks where the average is calculated over all possible locations of the break points, using the prior probability that there is a break point in each location as weight. We choose $g_i^j(\mathcal{O})$ to be uniform over each $(j, \mathcal{O})$ and set $\bar{\psi} << \sqrt{(N/2)}$.

Examining the hypothesis that the dynamics of the cross section are group-based, given $\mathcal{O}$, is equivalent to verifying the hypothesis that there are ψ breaks against the null of no breaks. Such an hypothesis can be examined with a Posterior odds ratio:

$$PO(\mathcal{O}) \quad = \quad \frac{g_0 f(Y|H_0)}{\sum_\psi g_\psi f^0(Y|H_\psi, \mathcal{O}) \mathbb{J}_1(n)} \tag{10.58}$$

where g_0 (g_ψ) is the prior probability that there are 0 (ψ) breaks. Verification of the hypothesis that there are $\psi - 1$ vs. ψ breaks in the cross section can be done using:

$$PO(\mathcal{O}, \psi - 1) \quad = \quad \frac{g_{\psi-1} f^{0(\psi-1)}(Y|H_{\psi-1}, \mathcal{O})}{g_\psi f^{0(\psi)}(Y|H_\psi, \mathcal{O}) \mathbb{J}_2(n)} \tag{10.59}$$

Here $\mathbb{J}_i(n),\ i = 1, 2$ are penalty functions which account for the fact that a model with ψ breaks is more densely parametrized than a model with a smaller number of breaks. Once the number of break points has been found (say, equal to $\hat{\psi}$), we assign units to groups so as to provide the highest total predictive density, i.e. compute $f^-(Y|H_{\hat{\psi}}, \mathcal{O})$. Since there

are $\mathcal{O}$ possible permutations of the cross section over which to search for groups the optimal permutation rule of units in the cross section is the one which achieves $f^{\dagger}(Y|H_{\hat{\psi}})$.

Two interesting questions which emerge are the following. Can we proceed sequentially to test for breaks? Bai (1997) shows that such a procedure produces consistent estimates of the number and the locations of the breaks. However, when there are multiple groups, the estimated break point is consistent for *any* of the existing break points and its location depends on the "strength" of the break. Second, how can we maximize the predictive density over $\mathcal{O}$ when n is large? When no information on the ordering of the units is available and n is moderately large, the approach is computationally demanding. Geographical, economic or sociopolitical factors may help to provide a restricted set of ordering worth examining. But even when economic theory is silent, the maximization does not require n! evaluations, since many orderings give the same predictive density.

Example 10.17 *Suppose n=4, so there are n!=24 possible orderings to examine. Suppose the initial ordering is 1234 and two groups are found: 1 and 234. Then all permutations of 234 with unit 1 coming ahead, i.e. 1243, 1342, etc., give the same predictive density. Similarly permutations which leave unit 1 last need not be examined, i.e. 2341, 2431, etc. This reduces the number of ordering to be examined to 13. By trying another ordering, say 4213, and finding, for example, two groups: 42 and 13, we can further eliminate all the orderings which rotate the elements of each group, i.e. 4132, 2341, etc.. It is easy to verify that once four carefully selected ordering have been tried and, say, two groups found in each trial, we have exhausted all possible combinations.*

Once the submodel characteristics have been determined, we can estimate $[\bar{\alpha}', \ vech(\Sigma_E)', \ vech(\Sigma_V)']'$ using $f^{\dagger}(Y|H_{\psi})$. For example, if e_{it}'s and v_i are normally distributed,

$$\begin{aligned}
\hat{\bar{\alpha}}^j &= \frac{1}{n^j(\mathcal{O})} \sum_{i=1}^{n^j(\mathcal{O})} \alpha^j_{i,ols} \\
\hat{\Sigma}_j &= \frac{1}{n^j(\mathcal{O})-1} \sum_{i=1}^{n^j(\mathcal{O})} (\alpha^j_{i,ols} - \hat{\bar{\alpha}}^j)(\alpha^j_{i,ols} - \hat{\bar{\alpha}}^j)' - \frac{1}{n^j(\mathcal{O})} \sum_{i=1}^{n^j(\mathcal{O})} (x_i x_i')^{-1} \hat{\sigma}_i^2 \\
\hat{\sigma}_i^2 &= \frac{1}{T-k}(y_i'y_i - y_i'x_i\alpha_{i,ols}) \qquad (10.60)
\end{aligned}$$

where x_i is the matrix of regressors and y_i the vector of dependent variables for unit i and $\alpha^j_{i,ols}$ is the OLS estimator of α^j obtained using the information for unit i (in group $j = 1, \ldots, \psi+1$). Then an Empirical Bayes posterior point estimate for the α vector is $\tilde{\alpha} = (X'\hat{\Sigma}_E^{-1}X + \hat{\Sigma}_V^{-1})^{-1}(X'\hat{\Sigma}_E^{-1}Y + \hat{\Sigma}_V^{-1}A\hat{\bar{\alpha}})$. Alternatively, if the e_{it}'s and the v_i's are normal and $g(a_0, \Sigma_E, \Sigma_V)$ is diffuse, we can jointly estimate $(\bar{\alpha}^j, \Sigma_j, \sigma_i^2)$ and the posterior mean for α as follows:

$$\hat{\bar{\alpha}}^j = \frac{1}{n^j(\mathcal{O})} \sum_{i=1}^{n^j(\mathcal{O})} (\alpha_i^*)^j$$

$$\begin{aligned}
\hat{\Sigma}_j &= \frac{1}{n^j(\mathcal{O}) - k - 1}[\delta * I + \sum_{i=1}^{n^j(\mathcal{O})} ((\alpha_i^*)^j - \hat{\bar{\alpha}}^j)((\alpha_i^*)^j - \hat{\bar{\alpha}}^j)'] \\
\hat{\sigma}_i^2 &= \frac{1}{T+2}(y_i - x_i\alpha_i^*)'(y_i - x_i\alpha_i^*) \\
(\alpha_i^*)^j &= (\frac{1}{\hat{\sigma}_i^2}x_i'x_i + \hat{\Sigma}_j^{-1})^{-1}(\frac{1}{\hat{\sigma}_i^2}x_i'x_i\alpha_{i,ols} + \hat{\Sigma}_j^{-1}\hat{\bar{\alpha}}^j) \qquad (10.61)
\end{aligned}$$

$j = 1, \ldots, \psi + 1; i = 1, \ldots, n^j(\mathcal{O})$; and $\delta > 0$ but small insures that $\hat{\Sigma}_j$ is positive definite.

Exercise 10.44 *Derive (10.60) and (10.61).*

Example 10.18 *(Convergence clubs). The cross sectional posterior distribution of steady states in example 10.15 shows a multimodal shape. One may therefore be interested in knowing whether there are convergence clubs in the data and where the break point is.*

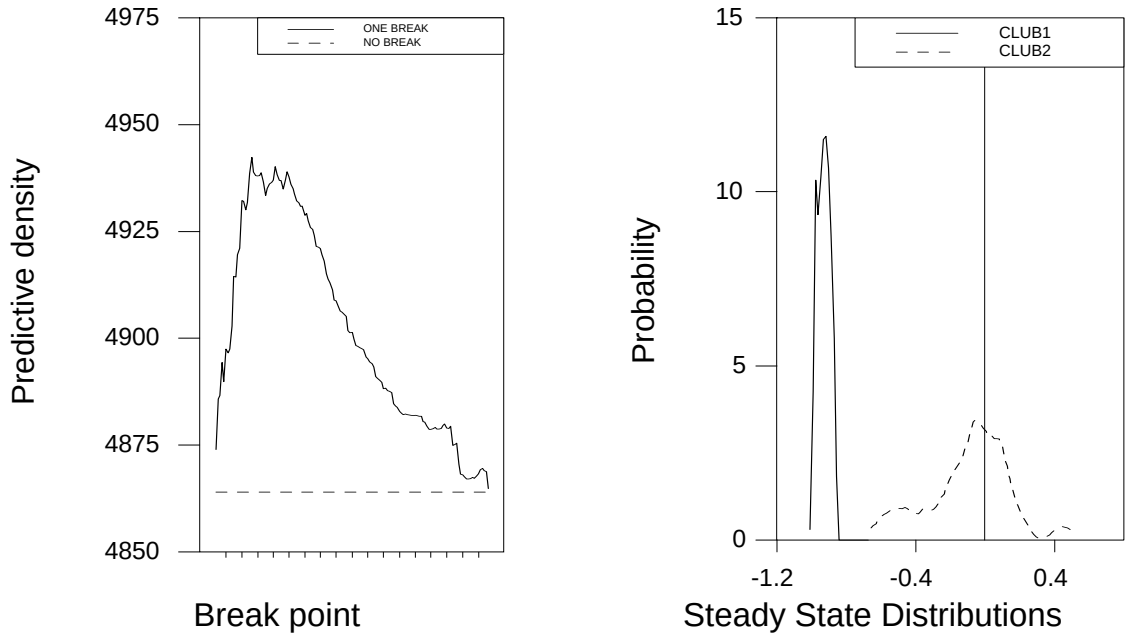


Figure 10.5: Convergence clubs.

We examined several ordering of cross sectional units based on initial income conditions, growth patterns or geographical characteristics. The one which is optimal orders units using the initial conditions of relative income per-capita. With this ordering, we set $\bar{\psi} = 4$ and sequentially examine ψ against $\psi + 1$ breaks starting from $\psi = 0$. There are up to three breaks in the data with PO ratios of 0.06. 0.52, 0.66 respectively. Conditioning on one

break ($\psi = 1$) we plot in the first panel of figure 10.5 the marginal predictive density as a function of the break point, together with the predictive density for $\psi = 0$. Visual inspection indicates that the former is always above the latter and that units up to 23 belong to the first group and from 24 to 144 to the second. The average convergence rates of the two groups are 0.78 and 0.20, suggesting faster convergence to below- average steady states in the first group. The second panel of figure 10.5 suggests that the posterior distributions of the steady states for the two groups are distinct. Not surprisingly, the first 23 units are all poor, Mediterranean and peripheral regions of the EU.

10.5.3 Panel VARs with interdependencies

Neither the panel VAR model studied in chapter 8 nor the specification we have considered so far allow for cross units lagged feedbacks. This may be important e.g. when one is interested in the transmission of shocks across countries. A panel VAR model with interdependencies has the form:

$$y_{it} = A_{1it}(\ell)y_t + A_{2it}(\ell)Y_t + e_{it} \tag{10.62}$$

where $i = 1, ..., n;\ t = 1, ..., T;\ y_{it}$ is a $m_1 \times 1$ vector for each i, $y_t = (y'_{1t}, y'_{2t}, \dots y'_{nt})'$, A^j_{1it} are $m_1 \times (nm_1)$ matrices and A^j_{2it} are $m_1 \times m_2$ matrices for each j; Y_t is a $m_2 \times 1$ vector of exogenous variables, common to all i, e_{it} is a $m_1 \times 1$ vector of disturbances and, for convenience, we have omitted constants and other deterministic components. In (10.62) cross-unit lagged interdependencies appear whenever $A^j_{1it,i'} \neq 0$, for $i' \neq i$ and some j, that is, when the matrix of lagged coefficients is not block diagonal at all lags. The presence of lagged cross unit interdependencies adds flexibility to the specification but it is not costless: the number of coefficients is greatly increased (there are $k = nm_1q_1 + m_2q_2$ coefficients in each equation). In (10.62) we allow coefficients to vary over time.

To construct posterior distributions for the unknowns, rewrite (10.62) as:

$$Y_t = X_t\alpha_t + E_t \qquad E_t \sim \mathbb{N}(0, \Sigma_E) \tag{10.63}$$

where $X_t = (I_{nm} \otimes \mathsf{X}_t)$; X_t=$(y'_{t-1},\ y'_{t-2}, \dots, y'_{t-q_1},\ Y'_t, \dots, Y'_{t-q_2})$; $\alpha_t = (\alpha'_{1t}, \dots, \alpha'_{nt})'$ and $\alpha_{it} = (\alpha^{1\prime}_{it}, \dots, \alpha^{m_1\prime}_{it})'$. Here α^j_{it} are $k \times 1$ vectors containing the coefficients for equation j of unit i, while Y_t and E_t are $nm \times 1$ vectors containing the endogenous variables and the random disturbances.

Whenever α_t varies with cross–sectional units in different time periods, it is impossible to employ classical methods to estimate it. Two short cuts are typically used: either it is assumed that the coefficient vector does not depend on the unit (apart from a time invariant fixed effect), or that there are no interdependencies (see e.g. Holtz Eakin et al. (1988) or Binder et al (2001)). Neither of these assumptions is appealing in our context. Instead, we assume that α_t can be factored as:

$$\alpha_t = \Xi_1\theta^1_t + \Xi_2\theta^2_t + \sum_{f=3}^{F} \Xi_f\theta^f_t \tag{10.64}$$

where Ξ_1 is a vector of ones of dimensions $nmk \times 1$; Ξ_2 is a matrix of ones and zeros of dimensions $nmk \times n$, and Ξ_f are conformable matrices. Here θ_t^2 is an $n \times 1$ vector of unit specific factors (the fixed effect), θ_t^1 is the common factor and θ_t^f is a set of factors which, in principle, is indexed by the unit i, the variable j, the lag or combinations of all of the above.

Example 10.19 *In a two variable, two lag, two country model with $Y_t = 0$, (10.64) implies*

$$\alpha_t^{i,j,s,\ell} = \theta_t^1 + \theta_t^{2i} + \theta_t^{3j} + \theta_t^{4s} + \theta_t^{5\ell} \tag{10.65}$$

where θ_t^1 is a common factor, $\theta_t^2 = (\theta_t^{21},\ \theta_t^{22})'$ is a 2×1 vector of country specific factors, $\theta_t^3 = (\theta_t^{31},\ \theta_t^{32})'$ is a 2×1 vector of equation specific factors, $\theta_t^{4s} = (\theta_t^{41},\ \theta_t^{42})'$ is a 2×1 vector of variable specific factors, $\theta_t^{5\ell} = (\theta_t^{51},\ \theta_t^{52})'$ is a 2×1 vector of lag specific factors.

All factors in (10.64) are allowed to be time varying; in fact, time invariant structures can be obtained via restrictions on the law of motion of the θ_t. Also, while the factorization in (10.64) is exact, in practice only a few factors will be specified: in that case all the omitted factors will be aggregated into an error term v_{1t}. Note also that with (10.64) the over-parametrization of the original model is dramatically reduced because the $nmk \times 1$ vector α_t depends on a much lower dimensional vector of factors.

Let $\theta_t = [\theta_t^1, (\theta_t^2)', (\theta_t^3)', \ldots, (\theta_t^{f_1})',\ f_1 < F]$ and write (10.64) as

$$\alpha_t = \Xi\theta_t + v_{1t} \quad v_{1t} \sim \mathbb{N}(0, \Sigma_E \otimes \Sigma_V) \tag{10.66}$$

where $\Xi = [\Xi_1, \Xi_2, \ldots, \Xi_{f_1}]$ and V is a $k \times k$ matrix. We assume a hierarchical structure on θ_t which allows for time variations and exchangeability:

$$\theta_t = (I - \mathbb{D}_1)\,\bar{\theta} + \mathbb{D}_1\theta_{t-1} + v_{2t} \qquad v_{2t} \sim \mathbb{N}\left(0, \Sigma_{v_{2t}}\right) \tag{10.67}$$

$$\bar{\theta} = \mathbb{D}_0\theta_0 + v_3 \qquad v_3 \sim \mathbb{N}(0, \Sigma_{v_3}) \tag{10.68}$$

We set $\Sigma_V = \sigma_v^2 I_k$ and, as in section 10.4, we let $\Sigma_{v_{2t}} = \phi_3 * \Sigma_{v_{2t-1}} + \phi_2 * \Sigma_0$ where $\Sigma_0 = diag(\Sigma_{01}, \Sigma_{02}, \ldots \Sigma_{0,f_1})$. We assume that $v_{it},\ i = 1, 2, 3$ and E_t are mutually independent and that $(\sigma_v^2, \phi_3, \phi_2, \mathbb{D}_1, \mathbb{D}_0)$ are known. Here $\mathbb{D}_0$ a matrix which restricts (part of the) means of the factors of the coefficients via an exchangeable prior.

To sum up, the prior for α_t has a multi-step hierarchical structure: with (10.66) we make a large number of coefficients depend on a smaller number of factors. The factors are then allowed to have a general evolving structure (equation (10.67)) and the prior mean of e.g. unit specific factors is potentially linked across units (equation (10.68)). The variance of the innovations in θ_t is allowed to be time varying to account for heteroschedasticity and other generic volatility clustering that are unit specific or common across units. To complete the specification we need to provide prior densities for $(\Sigma_E^{-1},\ \theta_0,\ \sigma_v^{-2},\ \Sigma_0^{-1},\ \Sigma_{v_3}^{-1})$. Canova and Ciccarelli (2002) study both informative and uninformative priors. Here we consider a special case of the non-informative framework they use.

Since α_t is a $nmk \times 1$ vector, the derivation of its posterior distribution with numerical methods is computationally demanding when m or n are large. To avoid problems rewrite the model as

$$\begin{aligned} y_t &= X_t \Xi \theta_t + e_t \\ \theta_t &= (I - \mathbb{D}_1)\,\bar{\theta} + \mathbb{D}_1 \theta_{t-1} + v_{2t} \\ \bar{\theta} &= \mathbb{D}_0 \theta_0 + v_{3t} \end{aligned} \tag{10.69}$$

where $e_t = E_t + X_t v_{1t}$ has covariance matrix $\sigma_t \Sigma_E = (1 + \sigma^2 X_t' X_t)\Sigma_E$. In (10.69) we have integrated α_t out of the model so that θ_t becomes the vector of parameters of interest.

We assume $\Sigma_{o1} = \phi_1,\ \Sigma_{0i} = \phi_i * I,\ i = 2, \ldots, f_1$, where ϕ_i controls the tightness of factor i of the coefficient vector. Furthermore assume that: $g(\Sigma_E^{-1}, \sigma^{-2}, \theta_0, \sigma_v^{-2}, \Sigma_{v_3}, \phi_i) = g(\Sigma_E^{-1}) g(\sigma^{-2}) g(\sigma_v^{-2}) g(\theta_0, \Sigma_{v_3}) \prod_i g(\phi_i)$ where $g(\Sigma_E^{-1})$ is $\mathbb{W}(\bar{\nu}_1, \bar{\Sigma}_1^{-1})$; $g(\sigma^{-2} \propto$ constant; $g(\sigma_v^{-2}) \propto \sigma_v^{-2}$; $g(\theta_0, \Sigma_{v_3}) \propto \Sigma_{v_3}^{-(\bar{\nu}_2+1)/2}$ where $\bar{\nu}_2 = 1 + N + \sum_{j=1}^{m_1} dim(\theta_{j,t}^f), f > 1$ and $g(\phi_i) \propto (\phi_i)^{-1}$; and the hyperparameters $\bar{\Sigma}_1, \bar{\nu}_1$ are assumed to be known or estimable from the data. The assumptions made imply that the prior for e_t has the form $(e_t | \sigma_t) \sim \mathbb{N}(0, \sigma_t \Sigma_E)$, and σ_t^{-2} is Gamma distributed so that e_t is distributed as a multivariate t centered at 0, with scale matrix which depends on Σ_E and degrees of freedom equal to $\dim(X_t)$. Since the likelihood of the data is proportional to $\left(\prod_{t=1}^T \sigma_i\right)^{-Nm/2} |\Sigma_E|^{-T/2} \exp\left[-\frac{1}{2} \sum_t (y_t - X_t \Xi \theta_t)' (\sigma_t \Sigma_E)^{-1} (y_t - X_t \Xi \theta_t)\right]$, it is easy to derive the conditional posteriors of the unknowns since the prior is conjugate. In fact, conditional on the other parameters, Σ_E^{-1} is Wishart, σ_t^{-2} is a Gamma, θ_0 is Normal, $\Sigma_{v_3}^{-1}$ is Wishart and ϕ_i^{-1} is Gamma distributed.

Exercise 10.45 *Derive the parameters of the posterior of Σ_E^{-1}, σ_t^{-1}, $\Sigma_{v_3}^{-1}$, ϕ_i^{-1} and θ_0.*

Finally, the conditional posterior distribution of $(\theta_1, .., \theta_T \mid y^T, \psi_{-\theta_t})$ can be obtained with the Kalman filter/ smoother as described in section 10.4. With these conditional, the Gibbs sampler can be used to draw a sequence of parameters from the joint posterior.

10.5.4 Indicators

The panel VAR (10.63) with the hierarchical prior (10.66)- (10.68) provides a framework to recursively construct coincident/leading indicators. In fact, the first equation in (10.69) is

$$y_t = \sum_{f=1}^{f_1} X_{f,t} \theta_t^f + e_t \tag{10.70}$$

where $X_{ft} = X_t \Xi_f$. In (10.70) y_t depends on a common time index X_{1t}, on a $n \times 1$ vector of unit specific indices X_{2t}, and of a set of indices which depend on variables, lags, units, etc. These indices are particular combinations of lags of the VAR variables, while θ_t^j measure the impact that different linear combinations of the lags of the right hand side variables have on the current endogenous variables. Hence, it is possible to construct leading indicators

directly from the VAR, without any preliminary distinction between leading, coincident and lagging variables. Also, because the model is recursive, both single-step and multi-step leading indicators can be obtained from the posterior for θ_t. Finally, fan charts can be constructed using the predictive density of future observations and the output of the Gibbs sampler.

Example 10.20 *Suppose we are interested in a model featuring a common, a unit specific and a variable specific indicator. Given (10.70), a leading indicator for y_t based on the common information available at time $t-1$ is $CLI_t = X_{1t}\theta^1_{t|t-1}$; a vector of leading indicators based on the common and unit specific information is $CULI_t = X_{1t}\theta^1_{t|t-1} + X_{2t}\theta^2_{t|t-1}$; a vector of indicators based on the common and variable specific information is $CVLI_t = X_{1t}\theta^1_{t|t-1} + X_{3t}\theta^3_{t|t-1}$; and vector of indicators based on the common, unit specific and variable specific information is $CUVLI_t = X_{1t}\theta^1_{t|t-1} + X_{2t}\theta^2_{t|t-1} + X_{3t}\theta^3_{t|t-1}$.*

While we have derived (10.70) using a prior on the panel VAR, one may want to start the investigation directly from (10.70). In this case, a researcher may be interested in assessing how many indices are necessary to capture the heterogeneities in the coefficients across time, units and variables. We can use Bayes factors to make this choice. A model with i indices is preferable to a model with $i+1$ indices, $i = 1, 2, \ldots, f_1 - 1$, if $\frac{f(y^{t+\tau}|\mathcal{M}_i)}{f(y^{t+\tau}|\mathcal{M}_{i+1})} > 1$ where $f(y^{t+\tau}|\mathcal{M}_i) = \int f(y^{t+\tau}|\theta_{t,i}, \mathcal{M}_i) g(\theta_{t,i}|\mathcal{M}_i) d\theta_{t,i}$ is the predictive density of a model with i indices for $y^{t+\tau} = [y_{t+1}, \ldots, y_{t+\tau}]$, $g(\theta_{t,i}|\mathcal{M}_i)$ is the prior for θ in model i and $f(Y^{t+\tau}|\theta_{t,i}, \mathcal{M}_i)$ the density of future data, given $\theta_{t,i}$ and $\mathcal{M}_i$. The predictive density for future $y_{t+\tau}$ in model i can be computed with the output of the Gibbs sampler. To do so, draw θ^l_t from the posterior distribution, construct forecast $y^l_{t+\tau}$ and prediction errors for each τ and average across draws.

10.5.5 Impulse responses

Impulse responses for the model can be computed as posterior revisions of the forecast errors. Since the model is non-linear, forecasts for the vector of endogenous variables may change because the innovations in the model or the innovations in the coefficients are different from zero. Furthermore, because of time variations, revisions depend on the history and the point in time where they are computed.

To see this set $Y_t = 0$, rewrite (10.63) as $\mathbb{Y}_t = \mathbb{A}_t \mathbb{Y}_{t-1} + \mathbb{E}_t$ and let $\alpha_t = vec(\mathbb{A}_{1t})$ where $\mathbb{A}_{1t}$ are the first m_1 rows of $\mathbb{A}_t$. Iterating τ times we have

$$y_{t+\tau} = \mathbb{S}(\prod_{s=0}^{\tau-1} \mathbb{A}_{t+\tau-s})\mathbb{Y}_t + \sum_{i=0}^{\tau-1} \mathbb{A}^*_{i,t+\tau}\, e_{t+\tau-i} \tag{10.71}$$

where $\mathbb{S} = [I, 0, .., 0]$ and $\mathbb{A}^*_{i,t+\tau} = \mathbb{S}(\prod_{s=0}^{i-1} \mathbb{A}_{t+\tau-s})\mathbb{S}'$; $\mathbb{A}^*_{0,t+\tau} = I$. Using (10.67) into (10.66) and iterating gives

$$\alpha_{t+\tau} = \Xi\theta_{t+\tau} + v_{1t+\tau} = \Xi\mathbb{D}_1^{\tau+1}\theta_{t-1} + \Xi\sum_{i=1}^{\tau} \mathbb{D}_1^i(I - \mathbb{D}_1)\bar{\theta} + \Xi\sum_{i=1}^{\tau} \mathbb{D}_1^i v_{2t+\tau-i} + v_{1t+\tau}$$

(10.72)

Define responses at step j, given information at t and terminal horizon τ as $Rev_{t,j}(\tau) = E_{t+j}\mathbb{Y}_{t+\tau} - E_t\mathbb{Y}_{t+\tau},\ \forall \tau \geq j+1$. Using $E_t y_{t+\tau} = \mathbb{S}E_t(\prod_{s=0}^{\tau-1} A_{t+\tau-s})\mathbb{Y}_t$, we have that

$$Rev_{t,j}(\tau) = \sum_{s=0}^{j-1}(E_{t+j}\mathbb{A}^*_{\tau-j+s,t+\tau})e_{t+j-s} + \mathbb{S}[E_{t+j}(\prod_{s=0}^{\tau-j-1}\mathbb{A}_{t+\tau-s})\prod_{s=\tau-j}^{\tau-1}\mathbb{A}_{t+\tau-s} - E_t(\prod_{s=0}^{\tau-1}\mathbb{A}_{t+\tau-s})]\mathbb{Y}_t \tag{10.73}$$

From (10.73) it is clear that forecast revisions can occur because new information present in the innovations of the model, e_t, or of the coefficients, v_{2t}, alter previous forecasts of $\mathbb{Y}_{t+\tau}$.

Example 10.21 *In equation (10.73) take $j = 1, \tau = 2$. Then $Rev_{t,1}(2) = E_{t+1}\mathbb{Y}_{t+2} - E_t\mathbb{Y}_{t+2} = E_{t+1}\left(\mathbb{A}^*_{1,t+2}\right)e_{t+1} + \mathbb{S}\left[E_{t+1}\left(\mathbb{A}_{t+2}\right)\mathbb{A}_{t+1} - E_t\left(\mathbb{A}_{t+2}\mathbb{A}_{t+1}\right)\right]\mathbb{Y}_t$. Similarly, $j = 2, k = 3$, imply $Rev_{t,2}(3) = E_{t+2}\mathbb{Y}_{t+3} - E_t\mathbb{Y}_{t+3} = \sum_{s=0}^{1}\left(E_{t+2}\mathbb{A}^*_{1+s,t+3}\right)e_{t+2-s} + \mathbb{S}[E_{t+2}\left(\mathbb{A}_{t+3}\right)\mathbb{A}_{t+2}\mathbb{A}_{t+1} - E_t\left(\mathbb{A}_{t+3}\mathbb{A}_{t+2}\mathbb{A}_{t+1}\right)]\mathbb{Y}_t$ where $\sum_{s=0}^{1}\left(E_{t+2}\mathbb{A}^*_{1+s,t+3}\right)e_{t+2-s} = \mathbb{S}E_{t+2}\left(\mathbb{A}_{t+3}\right)\mathbb{S}'e_{t+2} + \mathbb{S}E_{t+2}\left(\mathbb{A}_{t+3}\right)\mathbb{A}_{t+2}\mathbb{S}'e_{t+1}$. Hence, changes in $\mathbb{Y}_{t+3}$ due to innovations of the model are $\mathbb{S}E_{t+2}\left(\mathbb{A}_{t+3}\right)\mathbb{S}'e_{t+2} + \mathbb{S}E_{t+2}\left(\mathbb{A}_{t+3}\right)\mathbb{A}_{t+2}\mathbb{S}'e_{t+1}$ and due to innovations in the coefficients are $\mathbb{S}[E_{t+2}\left(\mathbb{A}_{t+3}\right)\mathbb{A}_{t+2}\mathbb{A}_{t+1} - E_t\left(\mathbb{A}_{t+3}\mathbb{A}_{t+2}\mathbb{A}_{t+1}\right)]\mathbb{Y}_t$. Clearly, responses depend on the time when they are generated (e.g. t vs. $t+1$) and the history of y_t.*

The output of the Gibbs sampler can be used to compute the expressions appearing in (10.73). Conditioning on $\mathbb{A}_t$, assuming that $e_t \neq 0$ and that all future innovations in both coefficients and variables are integrated out, $Rev_{t,j}(\tau)$ can be computed as follows:

Algorithm 10.4

1) *Draw $(e_{t+1}, \ldots, e_{t+j})$ and $(\mathbb{A}_{t+1}, \ldots, \mathbb{A}_{t+j})$ from the posterior distribution $L+1$ times.*

2) *For each draw $l = 2, \ldots L+1$, compute $\hat{A}^{*l}_{i,j} = \prod_{s=0}^{j}\mathbb{A}^l_{t+\tau-s}$. Average it $\hat{A}^{*l}_{i,j}$ over l.*

3) *For each draw $l = 2, \ldots, L+1$, compute $\hat{e}_{t+\tau} = \sum_{l=2}^{L+1} e^l_{t+\tau}$, $\tau > 1$.*

4) *Given $\mathbb{Y}_t$, $(e^l_{t+j}, \mathbb{A}^l_{t+j})$ from 1), $\hat{A}^{*l}_{i,j}$ from 2), $\hat{e}_{t+\tau}$ from 3), compute $Rev_{t,j}(\tau)$.*

Example 10.22 *We use a VAR model for G-7 countries with GDP growth, inflation, employment growth and the real exchange rate for each country and three indices: a 2×1 vector of common factors - one for EU and one for non-EU countries, a 7×1 vector of country specific factors and a 4×1 vector of variable specific factors.*

We assume time variations in the factors, use non-informative priors on the hyperparameters but do not impose exchangeability. Figure 10.6 presents 68% bands for the CUVLI indicator for EU GDP growth and inflation, constructed recursively using information available one year in advance. Actual values of EU GDP growth and inflation are superimposed. The model predicts the ups and downs of both series reasonably well using one year ahead information. The Theil-U statistics over the 1996:1-2000:4 and 1991:1-1995:4 sample are 0.87 and 0.66, respectively, much lower than those obtained with a single country VAR (1.25, 1.06) or with a univariate AR (1.04, 0.97).

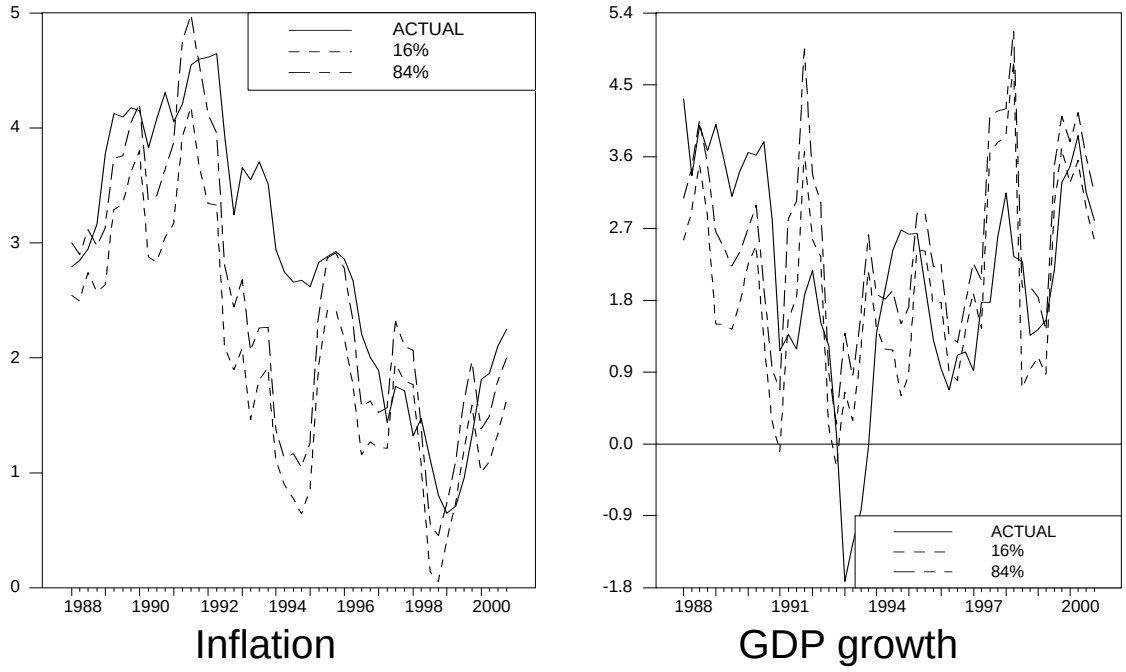


Figure 10.6: One year ahead 68% prediction bands, EU

Chapter 11: Bayesian time series and DSGE models

This chapter covers Bayesian estimation of three popular time series models and returns to the main goal of this book: estimation and inference in DSGE models, this time from a Bayesian perspective. All three types of time series models have a latent variable structure: the data y_t depends on an unobservable x_t and on a vector of parameters α, and the latent variable x_t is a function of another set of parameters θ. In factor models, x_t is a common factor or a common trend; in stochastic volatility models, x_t is a vector of volatilities and in Markov switching models, x_t is an unobservable finite-state process. While for the first and the third type of models classical methods to evaluate the likelihood function are available (see e.g. Sims and Sargent (1977), Hamilton (1989)), for the second approximations based on either a method of moments or quasi-ML are typically used. Approximations are needed because the density of observable data $f(y|\alpha,\theta)$ is a mixture of distributions i.e. $f(y|\alpha,\theta) = \int f(y|x,\alpha)f(x|\theta)dx$. Therefore, the computation of the likelihood function requires a T-dimensional integral and no analytical solution is available.

As mentioned in chapter 9, the model for x_t can be interpreted either as a prior or as a description of how the latent variable evolves. This means that all three models have an hierarchical structure which can be handled with the "data augmentation" technique of Tanner and Wong (1987). Such a technique treats $x^t = (x_1, \dots x_t)$ as a vector of parameters for which we have to compute the conditional posterior - as we have done with the time varying coefficients of a state space models in chapter 10. Cyclical sampling across the conditional distributions provides, in the limit, posterior draws for the parameters and the unobservable x^t. The Markov property for x_t is crucial to simplify the calculations since we can break the problem of simulating the x^t vector into the problem of simulating its components in a conditional recursive fashion. Since for this type of models the likelihood is bounded, if the priors are proper, the transition kernel induced by the Gibbs sampler (or by the mixed Gibbs-MH sampler) is irreducible, aperiodic and has an invariant distribution. Hence, sufficient conditions for convergence hold in these setups.

The kernel of (x,α,θ) is the product of the conditional distribution of $(y|x,\alpha)$, the conditional distribution of $(x|\theta)$ and the prior for (α,θ). Hence the marginal posterior for (α,θ), $g(\alpha,\theta|y) = \int g(x,\alpha,\theta|y)dx$, can be used for inference, while the marginal $g(x|y)$ provides a solution to the signal extraction problem. The main difference between this

setup and traditional signal extraction problems is that here we are interested in the whole distribution of x_t, not just its conditional mean. It is important to emphasize that, contrary to classical methods, the tools we describe allow the computation of the exact posterior distribution of the latent variable. Therefore, we are able to describe posterior uncertainty surrounding the latent variable and explicitly account for parameter uncertainty.

Forecasting y_t and the latent variable x_t is straightforward and can be handled with the tools described in chapter 9. Since many inferential exercises have to do with the problem of obtaining a future measure of the unobserved state (the business cycle in policy circles, the volatility process in business and finance circles), it is important to have ways to estimate it. Draws for future x_t can be obtained from the marginal posterior and the structure of the conditional of x_t.

Although this chapter primarily focuses on models with normal errors, more heavy-tailed distributions should be probably used, particularly in finance applications. As in the case of state space models, such an extension presents little complications.

The last section of the chapter goes back to DSGE models and studies how to obtain posterior estimates of the structural parameters, how to conduct posterior inference and model comparisons, and reexamines the link between DSGE and VARs. There is very little new material in this section: we bring together the models discussed in chapter 2, the ideas contained in chapters 5, 6 and 7 with the simulation techniques presented in chapter 9 to develop a framework where structural inference can be conducted in false models, taking both parameter and model uncertainty into consideration.

11.1 Factor Models

Factor models are used in many fields of economics and finance. They exploit the insight that there may be a source of common fluctuations in a vector of economic time series. Factor models are therefore alternatives to the (panel) VAR models analyzed in chapter 10. In the latter, detailed cross-variable interdependencies are modeled but no common factor is explicitly considered. Here, most of interdependencies are eschewed and a low dimensional vector of unobservable variables is assumed to drive the comovements across variables. Clearly, combinations of the two approaches are possible (see e.g. Bernanke, Boivin and Eliasz (2003), Uhlig (2003) or Giannone, Reichlin and Sala (2003)). The factor structure we consider is:

$$\begin{aligned} y_{it} &= \bar{y}_i + \mathbb{Q}_i y_{0t} + e_{it} \\ A_i^e(\ell) e_{it} &= v_{it} \\ A^y(\ell) y_{0t} &= v_{0t} \end{aligned} \tag{11.1}$$

where $E(v_{it}, v_{i't-\tau}) = 0,\ \forall i \neq i',\ i = 1, \ldots, m,\ E(v_{it}, v_{it-\tau}) = \sigma_i^2$, if $\tau = 0$ and zero otherwise, $E(v_{0t}, v_{0t-\tau}) = \sigma_0^2$ if $\tau = 0$ and zero otherwise, and y_{0t} is an unobservable factor. Two features of (11.1) need to be noted. First, the unobservable factor can have arbitrary serial correlation. Second, since the relationship between observables and unobservables is

static, e_{it} is allowed to be serially correlated. y_{0t} could be a scalar or a vector, as long as its dimensions is smaller than the dimension of y_t.

Example 11.1 *There are several specifications which fit into this framework. For example, y_{0t} could be a coincident business cycle indicator which moves a vector of macroeconomic time series y_{it}. In this case e_{it} captures idiosyncratic movements in y_{it}. Alternatively, y_{0t} could be a common stochastic trend while e_{it} is assumed to be stationary for all i. In this latter case (11.1) resembles the common trend-idiosyncratic cycle decomposition studied e.g. in Stock and Watson (1987). Furthermore, many of the models used in finance have a structure similar to (11.1). For example, in a Capital Asset Pricing Model (CAPM), y_{0t} represents an unobservable market portfolio. An interesting case emerges when $e_t = (e_{1t}, \ldots, e_{mt})'$ follows a VAR, i.e. $A^e(\ell)e_t = v_t$, and $A^e(\ell)$ is of order q_e, $\forall i$.*

We need restrictions to identify the parameters of (11.1). Since $\mathbb{Q}_i$ and y_{0t} are non-observable, both the scale and the sign of the factor and its loading cannot be separately identified. For normalization, we choose $\mathbb{Q}_{i1} > 0$ and assume that σ_0^2 is a fixed constant.

Let $\alpha_{1i} = (\bar{y}_i, \mathbb{Q}_i), i = 1, \ldots, m$ and $\alpha = (\alpha_{1i}, \sigma_i^2, A_{ij}^e,\ j = 1, \ldots, q_i, A_j^y, j = 1, \ldots q_0)$ be the vector of parameters of the model. Let $y_i = (y_{i1}, \ldots, y_{it})',\ y = (y_1', \ldots, y_m')'$. Given $g(\alpha)$, $g(\alpha|y, y_0) \propto f(y|\alpha, y_0)g(\alpha)$ and $g(y_0|\alpha, y) \propto f(y|\alpha, y_0)f(y_0|\alpha)$. To compute these conditional distributions, we need to derive $f(y|\alpha, y_0)$ and $f(y_0|\alpha) = \int f(y, y_0|\alpha)dy$.

Consider first $f(y|\alpha, y_0)$. Let $y_i^1 = (y_{i,1}, \ldots, y_{i,q_i})'$ be random and let $y_0^1 = (y_{0,1}, \ldots y_{0,q_0})'$ be the vector of initial observations on the factors, y_0^1 given, $A_i^e = (A_{i,1}^e, \ldots, A_{i,q_i}^e)$, $x_i^1 = [\mathbf{1}, y_0^1]$, where $\mathbf{1} = [1, 1, \ldots, 1]'$ and let $\mathbb{A}_i$ be a $(q_i \times q_i)$ companion matrix representation of $A_i^e(\ell)$. If the errors are normal, $(y_i^1|\bar{y}_i, \mathbb{Q}_i, \sigma_i^2, y_0^1) \sim \mathbb{N}(\bar{y}_i + \mathbb{Q}_i y_0^1, \sigma_i^2\Sigma_i)$, where Σ_i solves $\Sigma_i = \mathbb{A}_i\Sigma_i\mathbb{A}_i + (1, 0, \ldots 0)'(1, 0, \ldots 0)$.

Exercise 11.1 *Provide a closed form solution for Σ_i.*

Define $y_i^{1*} = \Sigma_i^{-0.5}y_i^1; x_i^{1*} = \Sigma_i^{-0.5}x_i^1$. To build the rest of the likelihood, let $e_i = [e_{i,q_i+1}, \ldots, e_{i,T}]'$ (this is $(T - q_i) \times 1$ vector); $e_{it} = y_{it} - \bar{y}_i - \mathbb{Q}_i y_{0t}$ and $E = [e_1, \ldots, e_{q_i}]$ (this is a $(T - q_i) \times q_i$ matrix). Similarly, let $y_0 = (y_{01}, \ldots, y_{0t})'$ and $Y_0 = (y_{0,-1}, \ldots, y_{0,-q_0})$. Let y_i^{2*} be a $(T - q_i) \times 1$ vector with the t-row equal to $A_i^e(\ell)y_{it}$ and let x_i^{2*} be a $(T - q_i) \times 2$ matrix with the t-row equal to $(A_i^e(1), A_i^e(\ell)y_{0t})$. Let $x_i^* = [x_i^{1*}, x_i^{2*}]',\ y_i^* = [y_i^{1*}, y_i^{2*}]$.

Exercise 11.2 *Write down the likelihood of $(y_i^*|x_i^*, \alpha)$ when e_t are normally distributed.*

To obtain $g(\alpha|y, y_0)$, assume that $g(\alpha) = \prod_j g(\alpha_j)$ and that $\alpha_{1i} \sim \mathbb{N}(\bar{\alpha}_{1i}, \bar{\Sigma}_{\alpha_{1i}}); A_i^e \sim \mathbb{N}(\bar{A}_i^e, \bar{\Sigma}_{A_i^e})\mathcal{I}_{[-1,1]};\ A^y \sim \mathbb{N}(\bar{A}^y, \bar{\Sigma}_{A^y})\mathcal{I}_{[-1,1]};\ \sigma_i^{-2} \sim \mathbb{G}(a_{1i}, a_{2i})$ where $\mathcal{I}_{[-1,1]}$ is an indicator function for stationarity, that is, the prior for $A_i^e(A^y)$ is normal, truncated outside the range $(-1, 1)$. The conditional posteriors are:

$$\begin{aligned}
(\alpha_{1i}|y_i,\ \alpha_{-\alpha_{1i}}) &\sim \mathbb{N}(\tilde{\Sigma}_{\alpha_{1i}}(\bar{\Sigma}_{\alpha_{1i}}^{-1}\bar{\alpha}_{1i} + \sigma_i^{-2}x_i^*y_i^*), \tilde{\Sigma}_{\alpha_{1i}}) \\
(A_i^e|y_i,\ y_0,\ \alpha_{-A_i^e}) &\sim \mathbb{N}(\tilde{\Sigma}_{A_i^e}(\bar{\Sigma}_{A_i^e}^{-1}\bar{A}_i^e + \sigma^{-2}E_i'e_i), \tilde{\Sigma}_{A_i^e})\mathcal{I}_{[-1,1]} \times \mathcal{N}(A_i^e) \\
(A^y|y_i,\ y_0,\ \alpha_{-A^y}) &\sim \mathbb{N}(\tilde{\Sigma}_{A^y}(\bar{\Sigma}_{A^y}^{-1}\bar{A}^y + \sigma^{-2}Y_0'y_0), \tilde{\Sigma}_{A^y})\mathcal{I}_{[-1,1]} \times \mathcal{N}(A^y) \\
(\sigma_i^{-2}|y_i,\ y_0,\ \alpha_{-\sigma_i}) &\sim \mathbb{G}(a_{1i} + T), (a_{2i} + (y_i^* - x_i^*\alpha_{1i,ols})^2)) \qquad (11.2)
\end{aligned}$$

where $\mathcal{N}(A_i^e) = |\Sigma_{A_i^e}|^{-0.5} \exp\{-\frac{1}{2\sigma^2}(y_i^1 - \bar{y}_i - \mathbb{Q}_i y_0^1)' \Sigma_{A_i^e}^{-1} (y_i^1 - \bar{y}_i - \mathbb{Q}_i y_0^1); \mathcal{N}(A^y) = |\Sigma_{A^y}|^{-0.5} \exp\{-\frac{1}{2\sigma^2}$ $(y_0^1 - A^y(\ell) y_{0,-1}^1)' \Sigma_{A_i^e}^{-1} (y_0^1 - A^y(\ell) y_{0,-1}^1)\}$; $\tilde{\Sigma}_{a_i} = (\bar{\Sigma}_{a_i}^{-1} + \sigma_i^{-2} x_i^{*\prime} x_i^*)^{-1}$; $\tilde{\Sigma}_{A_i^e} = (\bar{\Sigma}_{A_i^e}^{-1} +$ $\sigma^{-2} E_i' E_i)^{-1}$, $\tilde{\Sigma}_{A^y} = (\bar{\Sigma}_{A^y}^{-1} + \sigma^{-2} Y_0' Y_0)^{-1}$.

Sampling $(\bar{y}_i, \mathbb{Q}_i, \sigma_i^2)$ from (11.2) is straightforward. To impose the sign restriction necessary for identification, discard the draws producing $\mathbb{Q}_{i1} \leq 0$. The conditional posterior for $A_i^e(A^y)$ is complicated by the presence of the indicator for stationarity and the conditional distribution of the first $q_i(q_0)$ observations (absent these two, drawing these parameters would also be straightforward). Since these distributions are of unknown form, one could use the following variation of the MH algorithm to draw, e.g., A_i^e:

Algorithm 11.1

1) *Draw* $(A_i^e)^\ddagger$ *from* $\mathbb{N}(\tilde{\Sigma}_{A_i^e}(\bar{\Sigma}_{A_i^e}^{-1} \bar{A}^e{}_i + \sigma^{-2} E_i' e_i), \tilde{\Sigma}_{A_i^e})$. *If* $\sum_{j=1}^{q_i} (A_{i,j}^e)^\ddagger \geq 1$ *discard the draw.*

2) *Otherwise, draw* $\mathsf{U} \sim \mathbb{U}(0,1)$. *If* $\mathsf{U} < \mathcal{N}((A_i^e)^\ddagger)/\mathcal{N}((A_i^e)^{l-1})$, *set* $(A_i^e)^l = (A_i^e)^\ddagger$. *Else set* $(A_i^e)^l = (A_i^e)^{l-1}$.

3) *Repeat 1)-2) L times.*

The derivation of $g(y_0|\alpha, y)$ is a little more laborious but essentially straightforward. Define the $T \times T$ matrix $\mathcal{Q}_i^{-1} = \begin{bmatrix} \Sigma_i^{-0.5} & 0 \\ \Omega_i & \end{bmatrix}$) where Σ_i is a $q_i \times q_i$ matrix, 0 is a $q_i \times (T - q_i)$ matrix and $\Omega_i = \begin{bmatrix} -A_{i,q_i}^e & \dots & -A_{i,1}^e & 1 & 0 & \dots & 0 \\ 0 & -A_{i,q_i}^e & \dots & -A_{i,1}^e & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & -A_{i,q_i}^e & \dots & \dots & 1 \end{bmatrix}$). Similarly define $\mathcal{Q}_0^{-1}$. Let $x_i^\dagger = \mathcal{Q}_i^{-1} x_i$; $y_i^\dagger = \mathcal{Q}_i^{-1}(y_i - \mathbf{1}\bar{y}_i)$. Then the likelihood function is $\prod_{i=1}^m f(y_i^\dagger | \mathbb{Q}_i, \sigma_i^2, A_i^e, y_0)$ where $f(y_i^\dagger | \mathbb{Q}_i, \sigma_i^2, A_i^e, y_0) = (2\pi\sigma_i^2)^{-0.5T} \exp\{-(2\sigma_i^2)^{-1}(y_i^\dagger - \mathbb{Q}_i \mathcal{Q}_i^{-1} y_0)'(y_i^\dagger - \mathbb{Q}_i \mathcal{Q}_i^{-1} y_0)\}$. Since the marginal of the factor is $f(y_0|A^y) = (2\pi\sigma_0^2)^{-0.5T} \exp\{-(2\sigma_0^2)^{-1}(\mathcal{Q}_0^{-1} y_0)'(\mathcal{Q}_0^{-1} y_0)\}$, the joint likelihood of the data and the factor is $f(y^\dagger, y_0|\alpha) = \prod_{i=1}^m f(y_i^\dagger | \mathbb{Q}_i, \sigma_i^2, A_i^e, y_0) f(y_0|A^y)$. Completing the squares we have:

$$g(y_0 | y_i^\dagger, \alpha) \sim \mathbb{N}(\tilde{y}_0, \tilde{\Sigma}_{y_0}) \tag{11.3}$$

where $\tilde{y}_0 = \tilde{\Sigma}_{y_0}[\sum_{i=1}^m \mathbb{Q}_i \sigma_i^{-2} \mathcal{Q}_i^{-1'} \mathcal{Q}_i^{-1}(y_i - \mathbf{1}\bar{y}_i)]$; $\tilde{\Sigma}_{y_0} = (\sum_{i=0}^m \mathbb{Q}_i^2 \sigma_i^{-2} (\mathcal{Q}_i^{-1})'(\mathcal{Q}_i^{-1}))^{-1}$ with $\mathbb{Q}_0 = 1$. Note that $\tilde{\Sigma}_{y_0}$ is a $T \times T$ matrix. Given (11.2) and (11.3), the Gibbs sampler can be used to compute the joint conditional posterior of α and of y_0, and their marginals.

To make the Gibbs sampler operative we need to select σ_0^2 and the parameters of the prior distributions. For example, σ_0^2 could be set to the average variance of the innovations in a AR(1) regression for each y_{it}. Since little information is typically available on the loadings and the autoregressive parameters, it is a good idea to set $\bar{\alpha}_{i1} = \bar{A}_i^e = 0$ and assume a large prior variance. Finally, it is a good idea to choose a relatively diffuse prior for σ_i^{-2}, for example, $\mathbb{G}(4, 0.001)$, a distribution without the third and fourth moments.

The calculation of the predictive density of y_{0t} is straightforward and it is left as an exercise for the reader. Note that when the factor is a common business cycle indicator, the construction of this quantity produces the density of a leading indicator.

Exercise 11.3 *Describe how to construct the predictive density of* $y_{0t+\tau},\ \tau = 1, 2, \ldots$.

Exercise 11.4 *Suppose* $i = 4$ *and let* $A_i^e(\ell)$ *be of first order. In addition, suppose* $\bar{y} = [0.5, 0.8, 0.4, 0.9]'$; $\mathbb{Q}_1 = [1, 2, 0.4, 0.6, 0.5]'$. *Let* $A^e = diag[0.8, 0.7, 0.6, 0.9]$, $A^y = [0.7, -0.3]$, $v_0 \sim \mathbb{N}(0, 5)$ *and* $v \sim \mathbb{N}(0, \begin{bmatrix} 3 & 0 & 0 & 0 \\ 0 & 4 & 0 & 0 \\ 0 & 0 & 9 & 0 \\ 0 & 0 & 0 & 6 \end{bmatrix})$. *Let the priors be:* $(\bar{y}_i, \mathbb{Q}_i) \sim \mathbb{N}(0, 10 * I_2),\ i = 1, 2, 3, 4;\ A^e \sim \mathbb{N}(0, I_4)\mathcal{I}_{[-1,1]}\ \ A^y \sim \mathbb{N}(0, I_2)\mathcal{I}_{[-1,1]}$ *and* $\sigma_i^{-2} \sim \mathbb{G}(4, 0.001)$ *where* $\mathcal{I}_{[-1,1]}$ *instructs us to drop values of such that* $\sum_j A_{ij}^e \geq 1$ *or* $\sum_j A_j^y \geq 1$*). Draw sequences from the posterior of* α *and construct an estimate of the posterior distribution of* y_0.

Exercise 11.5 *Let the prior for* $(\bar{y}_i, \mathbb{Q}_i, A_i^e, A^y, \sigma_i^{-2})$ *be non-informative. Show that the posterior mean estimator for* y_0 *is the same as the one obtained by running the Kalman filter on model (11.1).*

11.1.1 Arbitrage Pricing (APT) Models

Apart from the construction of business cycle or trend indicators, factor models are extensively used in finance (see e.g. Campbell, Lo, MacKinley (1997) for references). Here the unobservable factor is a vector of portfolio excess returns; a vector of macroeconomic variables or a vector of portfolio of real returns, typically restricted to span the mean-variance frontier. APT models are useful since economic theory imposes restrictions on nonlinear combinations of the parameters of these models.

For illustrative purposes, consider a version of an APT model were a vector of m asset returns y_t is related to a vector of k factors y_{0t} according to the linear relationship

$$y_t = \bar{y} + \mathbb{Q}_1 y_{0t} + e_t \tag{11.4}$$

where $E(y_0) = 0, E(y_0 y_0') = I$, $E(e|y_0) = 0, E(ee'|y_0) = \Sigma_e$; $\bar{y}$ is a vector of conditional mean returns, $\mathbb{Q}_1$ is a $m \times k$ matrix of loadings and both $\mathbb{Q}_1$ and y_{0t} are unknown. Traditionally, a model like (11.4) is estimated in two steps: in the first step either the factor loadings or the factors themselves are estimated (with a cross sectional regression). Then, taking the first step estimates as if they were the true ones, a second pass regression (typically, in time series) is used to estimate the other parameters (see e.g. Roll and Ross (1980)). Clearly, this approach suffers from error-in-variables problems which can lead to incorrect inference.

A number of authors, starting from Ross (1976), have shown that, as $m \to \infty$, absence of arbitrage opportunities implies that $\bar{y}_i \approx \phi_0 + \sum_{j=1}^k \mathbb{Q}_{1ij}\phi_j$, where ϕ_0 is the intercept of the pricing relationship (the so-called zero-beta rate) and ϕ_j is the risk premium on factor $\mathbb{Q}_{1ij},\ j = 1, 2, \ldots, k$. With the two-step procedure we have described, and treating the

estimates of $\mathbb{Q}_{1i}$ and of $\bar{y}$ as given, the restrictions imposed become linear and tests can be easily developed e.g. using restricted and unrestricted estimates of ϕ_j (see Campbell, Lo and McKinley (1997)).

One way to test (11.4) is to measure the pricing errors and check their size relative to the average returns (with large relative errors indicating an inappropriate specification). This measure is given by $\mathsf{S} = \frac{1}{m}\bar{y}'[I - \mathbb{Q}(\mathbb{Q}'\mathbb{Q})^{-1}\mathbb{Q}']\bar{y}$ where $\mathbb{Q} = (\mathsf{1}, \mathbb{Q}_1)$ and $\mathsf{1}$ is a vector of ones of dimension m. For fixed m, $\mathsf{S} \neq 0$, while as $m \to \infty$, $\mathsf{S} \to 0$. It is typically hard to compute the sampling distribution of S. Using MCMC methods, its exact posterior distribution can be easily obtained.

For identification we require that $k < \frac{m}{2}$. Letting A_1^k be a lower triangular matrix containing the Choleski transformation of the first k independent rows of $\mathbb{Q}_1$, we also want $\mathbb{Q}_{1ii}^k > 0,\ i = 1, \ldots, k$.

Exercise 11.6 *Show that $k < \frac{m}{2}$ and $\mathbb{Q}_{1ii}^k > 0,\ i = 1, \ldots, k$ are necessary for identification.*

Let $\alpha_{i1} = (\bar{y}_i, \mathbb{Q}_i)$ and notice that, since the factors capture common components, $\Sigma_e = \text{diag}\{\sigma_i^2\}$. Then $f(\alpha_{i1}|y_0, \sigma_i) \propto \exp\{-\frac{1}{2\sigma_i^2}(\alpha_{i1} - \alpha_{i1,ols})'x'x(\alpha_{i1} - \alpha_{i1,ols})\}$, where $x = (\mathsf{1}, y_0)$ is a $T \times (k+1)$ matrix and $\alpha_{i1,ols}$ is the OLS estimators of the coefficients in a regression of y_{it} on $(1, y_0)$. We want to compute $g(\alpha|y_{0t}, y_t)$ and $g(y_{0t}|\alpha, y_t)$, where $\alpha = (\alpha_{1i}, \sigma_i^2, i = 1, 2, \ldots)$. We assume independence across i and standard priors for α, i.e. $\mathbb{Q}_{1i} \sim \mathbb{N}(\bar{\mathbb{Q}}_{1i}, \bar{\sigma}_{Q_1}^2),\ \mathbb{Q}_{1ii} > 0, i = 1, \ldots, k$; $\mathbb{Q}_{1i} \sim \mathbb{N}(\bar{\mathbb{Q}}_{1i}, \bar{\omega}_{Q_1}^2), i = k+1, \ldots, m$; $\bar{s}_i^2\sigma_i^{-2} \sim \chi^2(\bar{\nu}_i)$; $\bar{y}_i \sim \mathbb{N}(\bar{y}_{i0}, \bar{\sigma}_{\bar{y}_i}^2)$ where $\bar{y}_{i0} = \phi_0 + \sum_j \bar{\mathbb{Q}}_{1ij}\phi_j$ and ϕ_i are constant, while the hyperparameters of all prior distributions are given. Note that we impose the theoretical restrictions directly -the prior distribution of $\bar{y}_i$ is conditional on the value of $\mathbb{Q}_1$ - and that by varying $\bar{\sigma}_{y_i}^2$ we can account for different degrees of credence in the ATP restrictions. The conditional posterior distributions for the parameters are easily obtained.

Exercise 11.7 *Show that*

- $g(\bar{y}_i|y_t, y_{0t}, \mathbb{Q}_1, \sigma_i^2) \sim \mathbb{N}(\tilde{\bar{y}}_i, \tilde{\sigma}_{\bar{y}_i}^2)$ *where* $\tilde{\bar{y}}_i = \frac{(\bar{\sigma}_{y_i}^2\bar{y}_{i,ols} + (\sigma_i^2/T)\bar{y}_{i0})}{(\sigma_i^2/T) + \bar{\sigma}_{y_i}^2}$; $\tilde{\sigma}_{\bar{y}_i}^2 = \frac{(\sigma_i^2\bar{\sigma}_{y_i}^2)/T}{\sigma_i^2/T + \bar{\sigma}_{y_i}^2}$, $\bar{y}_{i,ols} = \frac{1}{T}\sum_{t=1}^T (y_{it} - \sum_{j=1}^k \mathbb{Q}_{1j}y_{0tj})$.

- $g(\mathbb{Q}_{1i}|y_t, y_{0t}, \bar{y}_i, \sigma_i^2) \sim \mathbb{N}(\tilde{\mathbb{Q}}_{1i}, \tilde{\Sigma}_{Q_{1i}})$, *where* $\tilde{\Sigma}_{Q_{1i}} = (\bar{\sigma}_{\mathbb{Q}_{1i}}^{-2} + \sigma_i^{-2}x_i^{\dagger\prime}x_i^{\dagger})^{-1}$; $\tilde{\mathbb{Q}}_{1i} = \Sigma_{Q_{1i}}(\bar{\mathbb{Q}}_{1i}\bar{\sigma}_{\mathbb{Q}_1}^{-2} + x_i^{\dagger\prime}x_i^{\dagger}\mathbb{Q}_{1i,ols}\sigma_i^{-2}), i = 1, \ldots, k$ *and* $\tilde{\Sigma}_{Q_{1i}} = (\bar{\omega}_{\mathbb{Q}_1}^{-2} + \sigma_i^{-2}x_i^{\dagger\prime}x_i^{\dagger})^{-1}$; $\tilde{\mathbb{Q}}_{1i} = \Sigma_{Q_{1i}}(\bar{\mathbb{Q}}_{1i}\bar{\omega}_{\mathbb{Q}_1}^{-2} + x_i^{\dagger\prime}x_i^{\dagger}\mathbb{Q}_{1i,ols}\sigma_i^{-2}), i = k+1, \ldots, m$ *where $\mathbb{Q}_{1i,ols}$ is the OLS estimator of a regression of $(y_{it} - \bar{y}_0)$ on $y_{01}, \ldots, y_{0i-1}$ and $x_i^{\dagger}$ is the matrix x_i without the first row.*

- $(\tilde{s}^2\sigma_i^{-2}|y_t, y_{0t}, \mathbb{Q}_1, \bar{y}_i) \sim \chi^2(\tilde{\nu})$ *where* $\tilde{\nu} = \bar{\nu} + T$; $\tilde{s}_i^2 = \bar{\nu}\bar{s}_i^2 + (T - k - 1)\sum_t (y_{it} - \bar{y}_i - \sum_j \mathbb{Q}_{1j}y_{0tj})^2$.

The joint of the data and of the factor is:

$$\begin{bmatrix} y_{0t} \\ y_t \end{bmatrix} \sim \mathbb{N}\left[\begin{pmatrix} 0 \\ \bar{y} \end{pmatrix}, \begin{pmatrix} I & \mathbb{Q}_1' \\ \mathbb{Q}_1 & \mathbb{Q}_1\mathbb{Q}_1' + \Sigma_e \end{pmatrix}\right]$$

Using the properties of conditional normal distributions we have $g(y_{0t}|y_t, \alpha) \sim \mathbb{N}(\mathbb{Q}_1'(\mathbb{Q}_1'\mathbb{Q}_1 + \Sigma_e)^{-1}(y_t - \bar{y}); I - \mathbb{Q}_1'(\mathbb{Q}_1'\mathbb{Q}_1 + \Sigma_e)^{-1}\mathbb{Q}_1)$, with $(\mathbb{Q}_1'\mathbb{Q}_1 + \Sigma_e)^{-1} = \Sigma_e^{-1} - \Sigma_e^{-1}\mathbb{Q}_1(I + \mathbb{Q}_1'\Sigma_e^{-1}\mathbb{Q}_1)^{-1}\mathbb{Q}_1'\Sigma_e^{-1}$, where $(I + \mathbb{Q}_1'\Sigma_e^{-1}\mathbb{Q}_1)$ is a $k \times k$ matrix.

Exercise 11.8 *Suppose the prior for α is non-informative, i.e. $g(\alpha) \propto \prod_j \sigma_{\alpha_j}^{-2}$. Derive the conditional posteriors for $\bar{y}, \mathbb{Q}_1, \Sigma_e$ and y_{0t} under this prior.*

Exercise 11.9 *Using monthly returns data on the stocks listed in Eurostoxx 50 for the last 5 years, construct 5 portfolios with the quintiles of the returns. Using informative priors compute the posterior distribution of the pricing error in a APT model using one and two factors (averaging over portfolios). You may want to try two values for σ_{0i}^2, one large and one small. Report the 68% posterior interval for* S*. Do you reject the theory? What can you say about the posterior mean of the proportion of idiosyncratic to total risk?*

11.1.2 Conditional Capital Asset Pricing models (CAPM)

A conditional capital asset pricing model combines data-based and model-based approaches to portfolio selection into a model of the form

$$\begin{aligned} y_{it+1} &= \bar{y}_{it} + \mathbb{Q}_{it} y_{0t+1} + e_{it+1} \\ \mathbb{Q}_{it} &= x_{1t}\phi_{1i} + v_{1it} \\ \bar{y}_{it} &= x_{1t}\phi_{2i} + v_{2it} \\ y_{0t+1} &= x_{2t}\phi_0 + v_{0t+1} \end{aligned} \tag{11.5}$$

where $x_t = (x_{1t}, x_{2t})$ is a set of observable variables, $e_{it+1} \sim \mathbb{N}(0, \sigma_e^2)$; $v_{0t+1} \sim \mathbb{N}(0, \sigma_0^2)$ and both v_{1it} and v_{2it} are assumed to be serially correlated, to take into account the possible misspecification of the conditioning variables x_{1t}. Here y_{it+1} is the return on asset i, y_{0t+1} is the return on an unobservable market portfolio. (11.5) fits the factor model structure we have so far considered when $v_{2it} = v_{1it} = 0$, $\forall t$, x_{2t} are the lags of y_{0t} and $x_{1t} = I$ for all t. Various versions of (11.5) have been considered in the literature.

Example 11.2 *Consider the model*

$$\begin{aligned} y_{it+1} &= \mathbb{Q}_{it} + e_{it+1} \\ \mathbb{Q}_{it} &= x_t\phi_i + v_{it} \end{aligned} \tag{11.6}$$

Here the return on asset i depends on an unobservable risk premium $\mathbb{Q}_{it}$ and on idiosyncratic error term and the risk premium is a function of observable variables.

If we relax the constant cost of risk assumption and allow time variations in the conditional variance of asset i, we have

$$\begin{aligned} y_{it+1} &= x_t\mathbb{Q}_t + e_{it+1} \quad & e_{it} \sim \mathbb{N}(0, \sigma_{e_i}^2) \\ \mathbb{Q}_t &= \mathbb{Q} + v_t \quad & v_t \sim \mathbb{N}(0, \sigma_v^2) \end{aligned} \tag{11.7}$$

Here the return on asset i depends on observable variables. The loading on the observables, assumed to be the same across assets, are allowed to vary over time. Note that by substituting the second expression into the first we have that the model's prediction error is heteroschedastic (the variance is $x_t' x_t \sigma_v^2 + \sigma_{e_i}^2$).

Exercise 11.10 *Suppose $v_{2it} = v_{1it} = 0$, $\forall t$ and assume that y_{0t} is known. Let $\alpha = [\phi_{21}, \ldots, \phi_{2m}, \phi_{11}, \ldots, \phi_{1m}]$. Assume a-priori that $\alpha \sim \mathbb{N}(\bar{\alpha}, \bar{\Sigma}_\alpha)$. Let the covariance matrix of $e_t = [e_{1t}, \ldots, e_{MT}]$ be Σ_e and assume that, a-priori, $\Sigma_e^{-1} \sim \mathbb{W}(\bar{\Sigma}^{-1}, \bar{\nu})$. Show that, conditional on $(y_{it}, y_{0t}, \Sigma_e, x_t)$, the posterior of α is normal with mean $\tilde{\alpha}$ and variance $\tilde{\Sigma}_\alpha$ and that the marginal posterior of Σ_e^{-1} is Wishart with scale matrix $(\bar{\Sigma}^{-1} + \Sigma_{ols}^{-1})$, where Σ_{ols} is the OLS estimate of the covariance matrix, and $\bar{\nu} + T$ degrees of freedom. Show the exact form of $\tilde{\alpha}$, $\tilde{\Sigma}_\alpha$ and Σ_{ols}.*

Exercise 11.11 *Still assume $v_{2it} = v_{1it} = 0$, $\forall t$ but allow y_{0t} to be unobservable. Postulate a law of motion for y_{0t} of the form $y_{0t+1} = x_{2t}\phi_0 + v_{0t+1}$, where x_{2t} are observables. Describe the steps needed to find the conditional posterior of y_{0t}.*

The specification in (11.5) is more complicated than the one we have examined in exercises 11.10-11.11 since time variation in the coefficients adds computational difficulties to the calculations of the posterior distribution. To highlight the steps involved, we describe a version of (11.5) where $v_{0it} = 0$ $\forall t$, $m = 1$, $x_t = x_{1t} = x_{2t}$ and we allow for AR(1) errors in the law of motion of $\mathbb{Q}_t$, that is:

$$\begin{aligned} y_{t+1} &= x_t\phi_2 + \mathbb{Q}_t y_{0t+1} + e_{t+1} \\ \mathbb{Q}_t &= (x_t - \rho x_{t-1})\phi_1 + \rho\mathbb{Q}_{t-1} + v_t \\ y_{0t} &= x_t\phi_0 + v_{0t} \end{aligned} \tag{11.8}$$

where ρ measures the persistence in $\mathbb{Q}_t$.

Let $\alpha = [\phi_0, \phi_1, \phi_2, \rho, \sigma_e^2, \sigma_v^2, \sigma_{v_0}^2]$ and let $g(\alpha) = \prod_j g(\alpha_j)$. Assume that $g(\phi_0) \sim \mathbb{N}(\bar{\phi}_0, \bar{\Sigma}_{\phi_0})$; $g(\phi_1) \sim \mathbb{N}(\bar{\phi}_1, \bar{\Sigma}_{\phi_1})$; $g(\phi_2) \sim \mathbb{N}(\bar{\phi}_2, \bar{\Sigma}_{\phi_2})$; $g(\rho) \sim \mathbb{N}(0, \bar{\Sigma}_\rho)\mathcal{I}_{[-1,1]}$; $g(\sigma_v^{-2}) \sim \chi(\bar{s}_v^2, \bar{\nu}_v)$; $g(\sigma_e^{-2},\ \sigma_{v_0}^{-2}) \propto \sigma_e^{-2}\sigma_{v_0}^{-2}$; and that all hyperparameters are known.

To construct the conditional posterior of $\mathbb{Q}_t$ note that, if ρ is known, $\mathbb{Q}_t$ can be easily simulated as in state space models. Therefore, partition $\alpha = (\alpha_1, \rho)$. Conditional on ρ we can rewrite the law of motion of $\mathbb{Q}_t$ as $y \equiv \mathbb{Q} - \rho\mathbb{Q}_{-1} = x^+\phi_1 + v$ where $\mathbb{Q} = [\mathbb{Q}_1, \ldots, \mathbb{Q}_t]'$, $x = [x_1, \ldots, x_t]'$, $x^+ = x - \rho x_{-1}$ and $v \sim \mathbb{N}(0, \sigma_v^2 I_T)$. Setting $\mathbb{Q}_{-1} = 0$ for $t = 0$, we have two sets of equations, one for the first observation and one for the others, i.e. $y_0 \equiv \mathbb{Q}_0 = x^+\phi_1 + v_0$ and $y_t \equiv \mathbb{Q}_t - \rho\mathbb{Q}_{t-1} = x_t^+\phi_1 + v_t$. The likelihood function $f(y|x, \phi_1, \rho)$ is proportional to $\propto (\sigma_v^2)^{-0.5T} exp\{-0.5[(y_0 - x_0^+\phi_1)\sigma_v^{-2}(y_0 - x_0^+\phi_1)' - \sum_{t=1}^T (y_t - x_t^+\phi_1)\sigma_v^{-2}(y_t - x_t^+\phi_1)']\}$.

Let $\phi_{1,ols}^0$ be the OLS estimator obtained from the first observation and $\phi_{1,ols}^1$ the OLS estimator obtained from the other observations. Combining the prior and the likelihood it is immediate to see that the posterior kernel of ϕ is proportional to $exp\{-0.5(\phi_1^0 - \phi_{1,ols}^0)' x_0^{'+} \sigma_v^{-2} x_0^+ (\phi_0^1 - \phi_{0,ols}^1) - 0.5 \sum_t (\phi_1^1 - \phi_{1,ols}^1)' x_t^{'+} \sigma_v^{-2} x_t^+ (\phi_1^1 - \phi_{1,ols}^1) - 0.5(\phi_1 - \bar{\phi}_1)' \bar{\Sigma}_{\phi_1}^{-1} (\phi_1 -$

$\bar{\phi}_1)\}$. Therefore, the conditional posterior for ϕ_1 is normal. The mean is a weighted average of prior mean and two OLS estimators, i.e. $\tilde{\phi}_1 = \tilde{\Sigma}_{\phi_1}(\bar{\Sigma}_{\phi_1}^{-1}\bar{\phi}_1 + x_0^{+\prime}\sigma_v^{-2}y_0 + \sum_t x_t^{+\prime}\sigma_v^{-2}y_t)$, and $\tilde{\Sigma}_{\phi_1} = (\bar{\Sigma}_{\phi_1}^{-1} + x_0^{+\prime}\sigma_v^{-2}x_0^{+} + \sum_t x_t^{+\prime}\sigma_v^{-2}x_t^{+})^{-1}$. The conditional posterior for σ_v^2 can be found using the same logic.

Exercise 11.12 *Show that the posterior kernel for σ_v^2 is $(\sigma_v^2)^{-0.5(T-1)}exp\{-0.5\sum_t \sigma_v^{-2}(y_t - x_t^{+}\phi_1)'(y_t - x_t^{+}\phi_1)\} \times (\sqrt{\frac{\sigma_v^2}{1-\phi_1^2}})^{(-0.5(\bar{\nu}_v+1+2))}exp\{-0.5(\frac{\sigma_v^2}{1-\phi_1^2})^{-1}((y_0 - x_0^{+}\phi_1)'(y_0 - x_0^{+}\phi_1) + \bar{\nu}_v)\}$. Suggest an algorithm to draw from this (unknown) distribution*

Once the distribution for the components of α_1 is found, we can use the Kalman filter/smoother to construct $\mathbb{Q}_t$ and the posterior of y_{0t}, conditional on ρ. To find the posterior distribution of ρ requires little more work. Conditional on ϕ_1, rewrite the law of motion for $\mathbb{Q}_t$ as $y_t^{\dagger} \equiv \mathbb{Q}_t - x_t\phi_1 = x_{t-1}^{\dagger}\rho + v_t$ where $x_{t-1}^{\dagger} = \mathbb{Q}_{t-1} - x_{t-1}\phi_1$. Once again, split the data in two: initial observations $(y_1^{\dagger}, x_0^{\dagger})$ and the rest $(y_t^{\dagger},\ x_{t-1}^{\dagger})$. The likelihood function is

$$\begin{aligned} f(y^{\dagger}|x^{\dagger},\phi_1,\rho) &\propto \sigma_v^{-T-1}exp\{-0.5(y_1^{\dagger} - x_0^{\dagger}\phi_1)'\sigma_v^{-2}(y_1^{\dagger} - x_0^{\dagger}\phi_1)\} \\ &+ exp\{-0.5(\sum_t y_t^{\dagger} - x_{t-1}^{\dagger}\phi_1)'\sigma_v^{-2}(y_t^{\dagger} - x_{t-1}^{\dagger}\phi_1)\} \end{aligned} \tag{11.9}$$

Let ρ_{ols} be the OLS estimator of ρ obtained with T data points. Combining the likelihood with the prior produces a kernel of the form $exp\{-0.5[\sum_t(\rho - \rho_{ols})'(x_t^{\dagger})'\sigma_v^{-2}x_t^{\dagger}(\rho - \rho_{ols}) + (\rho - \bar{\rho}')\bar{\Sigma}_{\rho}^{-1}(\rho - \bar{\rho})\} \times (\sqrt{\frac{\sigma_v^2}{1-\phi_1^2}})^{-0.5(\bar{\nu}_v+1+2)}exp\{-0.5(\frac{\sigma_v^2}{1-\phi_1^2})^{-1}\bar{\nu}_v + (y_1^{\dagger})'(\frac{\sigma_v^2}{1-\phi_1^2})^{-1}y_1^{\dagger}\}$. Hence, the conditional posterior for ρ is normal with mean $\tilde{\rho} = \tilde{\Sigma}_{\rho}(\bar{\Sigma}_{\rho}^{-1}\bar{\rho} + \sum_t(x_t^{\dagger})'\sigma_v^{-2}y_t^{\dagger})$, variance $\tilde{\Sigma}_{\rho} = (\bar{\Sigma}_{\rho}^{-1} + \sum_t x_t^{\dagger\prime}\sigma_v^{-2}x_t^{\dagger})^{-1}$, truncated outside the range $[-1, 1]$.

Exercise 11.13 *Provide a MH algorithm to draw from the conditional posterior of ρ.*

Once $g(\alpha_1|\rho, y_{0t}, y_t)$, $g(\rho|\alpha_1, y_{0t}, y_t)$, $g(y_{0t}|\alpha_1, \rho, y_t)$ are available, they can be inserted in a standard Gibbs sampler to find the joint posterior of the quantities of interest.

11.2 Stochastic Volatility Models

Stochastic volatility models are alternatives to GARCH or TVC models. In fact, as these models, they can account for time varying volatility and leptokurthosis but produce excess kurthosis without heteroschedasticity. Typically, the log of σ_t^2 is assumed to follow an AR process. Therefore, changes in y_t are driven by errors in the model for the observables or errors in the model for $\ln\sigma_t^2$. Such a feature adds flexibility to the specification and produces richer dynamics for the observables as compared to, e.g., GARCH type models, where the same random variable drives both observables and volatilities.

The most basic stochastic volatility specification is:

$$\begin{aligned} y_t &= \sigma_t e_t & e_t &\sim \mathbb{N}(0,1) \\ \ln(\sigma_t^2) &= \rho_0 + \rho_1\ln(\sigma_{t-1}^2) + \sigma_v v_t & v_t &\sim\ iid\ \mathbb{N}(0,1) \end{aligned} \tag{11.10}$$

where v_t and e_t are independent. In (11.10) we have assumed, for simplicity, that y_t is demeaned. Hence, this specification could be used to model, e.g., asset returns or changes in exchange rates. Also, again for simplicity, only one lag of σ_t^2 is considered.

Let $y^t = (y_1, \ldots, y_t)$, $\sigma^t = (\sigma_1^2, \ldots, \sigma_t^2)$ and let $f(\sigma^t|\rho, \sigma_v)$, be the probability mechanism generating σ^t, where $\rho = (\rho_0, \rho_1)$. The density of the data is $f(y^t|\rho, \sigma_v) = \int f(y^t|\sigma^t) f(\sigma^t|\rho, \sigma_v)$ $d\sigma^t$. As in factor models, we treat σ^t as an unknown vector of parameters, whose conditional distribution needs to be found.

We postpone the derivation of the conditional distribution of (ρ, σ_v) to a later (more complicated) application and concentrate on the problem of drawing a sample from the conditional posterior of σ_t^2. First, notice that, because of the Markov structure, we can break the joint posterior of σ^t into the product of conditional posteriors of the form $g(\sigma_t^2|\sigma_{t-1}^2, \sigma_{t+1}^2, \rho, \sigma_v, y_t)$, $t = 1, \ldots T$. Second, these univariate densities have an unusual form: they are the product of a conditional normal for y_t and a log normal for σ_t^2

$$\begin{aligned} g(\sigma_t^2|\sigma_{t-1}^2, \sigma_{t+1}^2, \rho, \sigma_v, y_t) &\propto f(y_t|\sigma_t^2) f(\sigma_t^2|\sigma_{t-1}^2, \rho, \sigma_v) f(\sigma_{t+1}^2|\sigma_t^2, \rho, \sigma_v) \\ &\propto \frac{1}{\sigma_t} \exp\{-\frac{y_t^2}{2\sigma_t^2}\} \times \frac{1}{\sigma_t^2} \exp\{-\frac{(\log \sigma_t^2 - E_t(\sigma_t^2))^2}{2var(\sigma_t^2)}\} \quad (11.11) \end{aligned}$$

where $E_t(\sigma_t^2) = \frac{(\rho_0(1-\rho_1)+\rho_1(\ln \sigma_{t+1}^2 + \ln \sigma_{t-1}^2)}{1+\rho_1^2}$, $var(\sigma_t^2) = \frac{\sigma_v^2}{1+\rho_1^2}$. Because $g(\sigma_t^2|\sigma_{t-1}^2, \sigma_{t+1}^2, \rho, \sigma_v, y_t)$ is non-standard, we need either a candidate density to be used as importance sampling or an appropriate transition function to be used in a MH algorithm. There is an array of densities one could use as importance sampling densities. For example, Jacquier, Polson and Rossi (1994) noticed that the first term in (11.11) is the density of an inverse of Gamma distributed random variable, that is, $x^{-1} \sim \Gamma(a_1, a_2)$, while the second term can be approximated by an inverse of a Gamma distribution (matching first and second moments). The inverse of a Gamma is a good "blanketing" density for the log-normal because it dominates the latter on the right tail. Furthermore, the two parts of the posterior can be combined into one inverse Gamma with parameters $\tilde{a}_1 = \frac{(1-2\exp(var(\sigma_t^2)))}{1-\exp(var(\sigma_t^2))} + 0.5$ and $\tilde{a}_2 = [(\tilde{a}_1 - 1)(\exp(E_t(\sigma_t) + 0.5var(\sigma_t))) + 0.5y_t^2]$. Hence draws made from this target density. As an alternative, since the kernel of $\ln(\sigma_t^2)$ is of known form, we could draw $\ln(\sigma_t^2)$ from $\mathbb{N}(E(\sigma_t^2) - 0.5var(\sigma_t^2), var(\sigma_t^2))$ and accept the draw with probability equal to $\exp\{-0.5\frac{y_t^2}{\sigma_t^2}\}$ (see Geweke (1994)).

Example 11.3 *We have run a small Monte Carlo experiment to check the quality of these two approximations setting. Table 11.1 below reports the percentiles using 5000 draws from the posterior when* $\rho_0 = 0.0, \rho_1 = 0.8$ *and* $\sigma_v = 1.0$. *Both approximations appear to produce similar results.*

It is worthwhile stressing that (11.10) is a particular nonlinear Gaussian model which can be transformed into a linear but non-Gaussian state space model without loss of information. In fact, letting $x_t = \ln \sigma_t$; $\epsilon_t = \ln e_t^2 + 1.27$, equation (11.10) could be written as

$$\ln y_t^2 = -1.27 + x_t + \epsilon_t$$

	5th	*25th*	*median*	*75th*	*95th*
Gamma	*0.11*	*0.70*	*1.55*	*3.27*	*5.05*
Normal	*0.12*	*0.73*	*1.60*	*3.33*	*5.13*

Table 11.1: Percentiles of the approximating distributions

$$x_{t+1} = a + \rho x_t + \sigma_v v_t \tag{11.12}$$

where ϵ_t has zero mean but it is non-normal. A framework like this was encountered in chapter 10 and techniques designed to deal with such models were outlined there. Here it is sufficient to point out that a non-normal density for ϵ_t can be approximated with a mixture of normals, i.e. $f(\epsilon_t) \approx \sum_j \varrho_j f(\epsilon_t|\mathcal{M}_j)$, where $f(\epsilon_t|\mathcal{M}_j) \sim \mathbb{N}(\bar{\epsilon}_j, \sigma^2_{\epsilon_j})$, $j = 1, \dots J$, $0 \leq \varrho_j \leq 1$. Chib (1996) provides details on how this can be done.

Cogley and Sargent (2003) have recently applied the mechanics of stochastic volatility models to a BVAR with time varying coefficients. Since the setup is an alternative to the linear time-varying conditional structures we have studied in chapter 10, we will examine in details how to obtain conditional posterior estimates for the parameters of such a model.

A VAR model with stochastic volatility has the form:

$$\begin{aligned} y_t &= (I_m \otimes \mathsf{X}_t)\alpha_t + e_t \qquad e_t \sim \mathbb{N}(0, \Sigma_t^\dagger) \\ \Sigma_t^\dagger &= \mathcal{P}^{-1}\Sigma_t \mathcal{P}^{-1\prime} \\ \alpha_t &= \mathbb{D}_1 \alpha_{t-1} + v_{1t} \qquad v_{1t} \sim \mathbb{N}(0, \Sigma_{v_1}) \end{aligned} \tag{11.13}$$

where $\mathcal{P}$ is a lower triangular matrix with ones on the main diagonal, $\Sigma_t = \text{diag}\{\sigma^2_{it}\}$, $\ln \sigma^2_{it} = \ln \sigma^2_{it-1} + \sigma_{v_{2i}} v_{2it}$, and $\mathbb{D}_1$ is such that α_t is a stationary process. In (11.13) the process for y_t has time varying coefficients and time varying variances. To compute conditional posteriors note that it is convenient to block together the α_t's, and the σ^2_t's and draw a whole sequence for these two vectors of random variables.

We make standard prior assumptions, that is: $\alpha_0 \sim \mathbb{N}(\bar{\alpha}, \bar{\Sigma}_a)$; $\Sigma^{-1}_{v_1} \sim \mathbb{W}(\bar{\Sigma}_{v_1}, \bar{\nu}_{v_1})$ where $\bar{\Sigma}_{v_1} \propto \bar{\Sigma}_a$, $\bar{\nu}_\varepsilon = dim(\alpha_0) + 1$, $\sigma^{-2}_{v_{2i}} \sim \mathbb{G}(a_1, a_2)$, $\ln \sigma_{i0} \sim \mathbb{N}(\ln \bar{\sigma}_i, \bar{\Sigma}_\sigma)$ and letting ϕ represent the non zero elements of $\mathcal{P}$, $\phi \sim \mathbb{N}(\bar{\phi}, \bar{\Sigma}_\phi)$.

Given these priors, the calculation of the conditional posterior for $(\alpha_t, \Sigma_{v_1}, \sigma_{v_{2i}})$ is straightforward. The conditional posterior for α_t can be obtained with a run of the Kalman filter/smoother as detailed in chapter 10; the conditional posterior for $\Sigma^{-1}_{v_1}$ is $\mathbb{W}(\bar{\Sigma}_{v_1} + (\sum_t v_{1t} v'_{1t}), \bar{\nu}_{v_1} + T)$, and that for $\sigma^{-2}_{v_{2i}}$ is $\mathbb{G}(a_1 + T, a_2 + \sum_t (\ln \sigma^2_{it} - \ln \sigma^2_{it-1})^2)$.

Example 11.4 *Suppose* $y_t = \alpha_t y_{t-1} + e_t$, $e_t \sim \mathbb{N}(0, \sigma^2_t)$, $\alpha_t = \rho \alpha_{t-1} + v_{1t}$, $v_{1t} \sim \mathbb{N}(0, \sigma^2_{v_1})$, $\ln \sigma^2_t = \ln \sigma^2_{t-1} + \sigma_{v_2} v_{2t}$, $v_{2t} \sim \mathbb{N}(0, 1)$. *Given* ρ, *the conditional posterior of* $(\sigma^{-2}_{v_1}, \sigma^{-2}_{v_2})$ *are Gamma with parameters* $(a_{v_1} + T, \bar{s^2}_{v_1} + (\sum_t v^2_{1t}))$ *and* $(a_{v_2} + T, \bar{b}_{v_2} + \sum_t (\ln \sigma^2_t - \ln \sigma^2_{t-1})^2)$, *respectively.*

Exercise 11.14 *Derive the conditional posteriors of* $(\rho, \sigma^{-2}_{v_1}, \sigma^{-2}_{v_2})$ *in example 11.4 when* ρ *is unknown and has prior* $\mathbb{N}(\bar{\rho}, \bar{\sigma}^2_\rho)\mathcal{I}_{[-1,1]}$, *where* $\mathcal{I}_{[-1,1]}$ *is an indicator for stationarity.*

To construct the conditional of ϕ note that if $\epsilon_t \sim (0, \Sigma_t)$, then $e_t \sim (0, \mathcal{P}\Sigma_t\mathcal{P}')$. Hence, if e_t is known, the free elements of $\mathcal{P}$ can be estimated, given (y_t, x_t, α_t). Since $\mathcal{P}$ is lower triangular, the m-th equation is

$$\sigma_{mt}^{-1}e_{mt} = \phi_{m1}(-\sigma_{mt}^{-1}e_{1t}) + \ldots + \phi_{m,m-1}(-\sigma_{mt}^{-1}e_{m-1t}) + (\sigma_{mt}^{-1}\epsilon_{mt}) \tag{11.14}$$

Hence, letting $E_{mt} = (-\sigma_{mt}^{-1}e_{1t}, \ldots, -\sigma_{mt-1}^{-1}e_{mt})$, $\varepsilon_{mt} = -\sigma_{mt}^{-1}\epsilon_{mt}$, it is easy to see that the conditional posterior for ϕ_i is normal with mean $\tilde{\phi}_i$, and variance $\tilde{\Sigma}_{\phi_i}$.

Exercise 11.15 *Show the form of $\tilde{\phi}_i$ and $\tilde{\Sigma}_{\phi_i}$.*

To draw σ_{it}^2 from its conditional distribution let $\sigma_{(-i)t}^2$ be the sequence of σ_t^2 except its i-th element and let $e^t = (e_1, \ldots, e_t)$. Then $g(\sigma_{it}^2|\sigma_{(-i)t}^2, \sigma_{\epsilon_i}, e^t) = g(\sigma_{it}^2|\sigma_{it-1}^2, \sigma_{it+1}^2, \sigma_{\epsilon_i}, e^t)$ which is given in (11.11). To draw from this distribution we could, for each i, choose as candidate distribution $\sigma_{it}^{-2}\exp\{-\frac{(\ln\sigma_{it}^2 - E_t(\sigma_{it}^2))^2}{2var(\sigma_t^2)}\}$ and accept or reject the draw with probability $\frac{(\sigma_{it}^{\ddagger})^{-1}\exp\{-\frac{e_{it}^2}{2(\sigma_{it}^2)^{\ddagger}}\}}{(\sigma_{it}^{\ell-1})^{-1}\exp\{-\frac{e_{it}^2}{2(\sigma_{it}^2)^{l-1}}\}}$, where $(\sigma_{it}^2)^{l-1}$ is the last draw and $(\sigma_{it}^2)^{\ddagger}$ is the candidate draw.

Exercise 11.16 *Suppose you are interested in predicting future values of y_t. Let $y^{t+\tau} = (y_{t+1}, \ldots y_{t+\tau})$. Show that, conditional on time t information:*

$$\begin{aligned} g(y^{t+\tau}|\alpha^t, \Sigma_t^{\dagger}, \Sigma_{v_1}, \phi, \sigma_{v_{2i}}, y^t) &= g(\alpha^{t+\tau}|\alpha^t, \Sigma_t^{\dagger}, \Sigma_{v_1}, \phi, \sigma_{v_{2i}}, y^t) \\ &\times\ g(\Sigma^{\dagger,t+\tau}|\alpha^{t+\tau}, \Sigma_t^{\dagger}, \Sigma_{v_1}, \phi, \sigma_{v_{2i}}, y^t) \\ &\times\ f(y^{t+\tau}|\alpha^{t+\tau}, \Sigma^{\dagger,t+\tau}, \Sigma_{v_1}, \phi, \sigma_{v_{2i}}, y^t) \end{aligned}$$

Describe how to sample (y_{t+1}, y_{t+2}) from this distribution. How would you construct a 68 percent prediction band ?

Stochastic volatility models are typically used to infer values for unobservable conditional volatilities, both within sample (smoothing) and out-of-sample (prediction). For example, option pricing formulas require estimates of conditional volatilities and event studies often relate specific occurrences to changes in volatility. Here we concentrate on the smoothing problem, i.e. on the computation of $g(\sigma_t^2|y^T)$. Once this is obtained, we can use its mean as an estimate of the smoothed variance. An analytic expression for this posterior density is not available but we can estimate it using $g(\sigma_t^2|y^T) = \int g(\sigma_t^2, |\alpha_t, y^T)g(\alpha_t|y^T)d\alpha_t$. Hence $g(\sigma_t^2|y^T)$ can be numerically obtained using the draws of σ_t^2 and α_t. Notice that this density directly accounts for parameter uncertainty.

Exercise 11.17 *i) Suppose the volatility model is $\ln\sigma_t^2 = \rho_0 + \rho(\ell)\ln\sigma_{t-1}^2 + \sigma_v v_t$, where $\rho(\ell)$ is unknown of order q. Show how to extend the Gibbs sampler to this case.*
ii) Assume a model of the form $\ln\sigma_t^2 = \rho_0 + \rho_1\ln\sigma_{t-1}^2 + \sigma_{v_t}v_t$ where $\sigma_{v_t} = f(x_t)$, x_t are observable variables and f is linear. Show how to extend the Gibbs sampler to this case

As with factor models, cycling through the conditionals of $(\Sigma_t^\dagger, \alpha_t, \sigma_{v_{2i}}, \Sigma_{v_1})$ with the Gibbs sampler produces, in the limit, a sample from the joint posterior.

Uhlig (1994) proposed an alternative specification for a stochastic volatility model which, together with a particular distribution of the innovations of the stochastic volatility term, produces closed form solutions for the posterior distribution of the parameters and of the unknown vector of volatilities. The approach treats some parameters in the stochastic volatility equation as fixed but has the advantage of producing recursive estimates of the quantities of interest.

Consider an m variable VAR(q) with stochastic volatility of the form:

$$\begin{aligned} Y_t &= AX_t + \mathcal{P}_t^{-1} e_t \quad e_t \sim \mathbb{N}(0, I) \\ \Sigma_{t+1} &= \frac{\mathcal{P}_t' v_t \mathcal{P}_t}{\rho} \qquad v_t \sim \text{Beta}((\nu + k)/2, 1/2) \end{aligned} \tag{11.15}$$

where X_t contains the lags of the endogenous and the exogenous variables, $\mathcal{P}_t$ is the upper Choleski factor of Σ_t, ν and ρ are (known) parameters, *Beta* denotes the m-variate Beta-distribution and k is the number of parameters in each equation.

To construct the posterior of the parameters of (11.15) we need a prior for (A, Σ_1). We assume $g_1(A, \Sigma_1) \propto g_0(A) g(A, \Sigma_1 | \bar{A}_0, \rho\bar{\Sigma}_A, \bar{\Sigma}_0, \bar{\nu})$, where $g_0(A)$ is a function restricting the prior for A (for example, to be stationary) and $g(\alpha, \Sigma_1 | \bar{A}_0, \rho\bar{\Sigma}_A, \bar{\Sigma}_0, \bar{\nu})$ is of Normal-Wishart form, i.e. $g(A|\Sigma_1) \sim \mathbb{N}(\bar{A}_0, \rho\bar{\Sigma}_A)$, $g(\Sigma_1^{-1}) \sim \mathbb{W}(\bar{\Sigma}_0, \bar{\nu})$, $\bar{A}_0, \bar{\Sigma}_0, \bar{\Sigma}_A, \bar{\nu}, \rho$ known.

Combining the likelihood of (11.15) with these priors and exploiting the fact that the Beta distribution conjugates with the Gamma distribution, we have that the posterior kernel for (A, Σ_{t+1}) is $\grave{g}_t(A, \Sigma_{t+1}) = \grave{g}_t(A) \grave{g}(A, \Sigma_{t+1} | \tilde{A}_t, \rho\tilde{\Sigma}_{At}, \tilde{\Sigma}_t, \nu)$ where $\grave{g}$ is of Normal-Wishart type, $\tilde{\Sigma}_{At} = \rho\tilde{\Sigma}_{At-1} + X_t X_t'$; $\tilde{A}_t = (\rho\tilde{A}_{t-1}\tilde{\Sigma}_{At-1} + Y_t X_t')\tilde{\Sigma}_t^{-1}$; $\tilde{\Sigma}_t = \rho\tilde{\Sigma}_{t-1} + \frac{\rho}{\nu} e_t (1 - X_t'\tilde{\Sigma}_{At}^{-1} X_t)\tilde{e}_t'$, $\tilde{e}_t = Y_t - \tilde{A}_{t-1} X_t$ and $\grave{g}_t(A) = \grave{g}_{t-1}(A) |(A - \tilde{A}_t)\tilde{\Sigma}_{At}(A - \tilde{A}_t)' + \frac{\nu}{\rho}\tilde{\Sigma}_t|^{-0.5}$.

Example 11.5 *Consider a univariate AR(1) version of (11.15) of the form*

$$y_t = \alpha y_{t-1} + \sigma_t^{-1} e_t \quad e_t \sim \mathbb{N}(0, 1) \tag{11.16}$$

$$\rho\sigma_{t+1}^2 = \sigma_t^2 v_t \qquad v_t \sim \text{Beta}((\nu + 1)/2, 1/2) \tag{11.17}$$

Let $g(\alpha, \sigma_1^2) \propto g_0(\alpha) g(\alpha, \sigma_1^2 | \bar{\alpha}_0, \rho\bar{\sigma}_{\alpha_0}^2, \bar{\sigma}_0^2, \bar{\nu})$ *where* $(\bar{\alpha}_0, \sigma_{\alpha_0}, \bar{\sigma}_0,$
$\bar{\nu})$ *are hyperparameters and assume that* $g(\alpha, \sigma_1^2 | \bar{\alpha}_0, \rho\bar{\sigma}_{\alpha_0}^2, \bar{\sigma}_0^2, \bar{\nu})$ *is of Normal-Inverted gamma type. Recursive posterior estimates of the parameters and of* $g_t(\alpha)$ *are* $\tilde{\sigma}_{\alpha,t}^2 = \rho\tilde{\sigma}_{\alpha,t-1}^2 + y_{t-1}^2$; $\tilde{\alpha}_t = \frac{(\rho\tilde{\alpha}_{t-1}\sigma_{\alpha,t-1}^2 + y_t y_{t-1})}{\sigma_{\alpha,t}^2}$; $\tilde{\sigma}_t^2 = \rho\tilde{\sigma}_{t-1}^2 + \frac{\rho}{\nu}\tilde{e}_t^2 (1 - \frac{y_{t-1}^2}{\sigma_{\alpha,t}^2})$; $\tilde{e}_t = y_t - \tilde{\alpha}_{t-1} y_{t-1}$; $g_t(\alpha) = g_{t-1}(\alpha)((\alpha - \tilde{\alpha}_t)^2 \sigma_{\alpha,t}^2 + \frac{\nu}{\rho}\sigma_t^2)^{-0.5}$. *Hence both* $\tilde{\sigma}_{\alpha_t}^2$ *and* $\tilde{\alpha}$ *are weighted averages with* ρ *measuring the memory of the process. Note that past values of* $\tilde{\alpha}$ *are weighted by the relative change in* $\tilde{\sigma}_{\alpha,t}^2$. *When* σ_t^2 *is constant,* $\tilde{\alpha}_t = \rho\tilde{\alpha}_{t-1} + \frac{y_t y_{t-1}}{\rho\sigma_{t-1}^2}$.

When $\rho = \frac{\nu}{\nu+1}$; $\frac{\nu}{\rho} = 1 - \rho$, *so that* $\tilde{\sigma}_t^2$ *is a weighted average of* $\tilde{\sigma}_{t-1}^2$ *and the information contained in the square of the recursive residuals, adjusted for the relative size of* y_t^2, *to*

the weighted sum of y_{t-1}^2 *up to* $t-1$. *Note also that* $E_{t-1}\sigma_t^2 = \frac{\sigma_{t-1}^2(\nu+1)}{\rho(\nu+2)}$. *Hence, when* $\rho = \frac{\nu+1}{\nu+2}$, σ_t^2 *is a random walk. When* $\nu \to \infty$, $\sigma_1^2 = \bar{\sigma}_0^{-2}$, *and* $\sigma_t^2 = \frac{\sigma_{t-1}^2(\nu+1)}{\rho(\nu+2)}$.

For comparison, it may be useful to map the prior into a Minnesota-type prior. For example, we could set $\bar{\Sigma}_0 = diag\{\bar{\sigma}_{0i}\}$ and compute $\bar{\sigma}_{0i}$ from the average square residuals of an AR(1) regression for each i in a training sample. Also, one could set $\bar{\Sigma}_A$ =block diag $[\bar{\Sigma}_{A1}, \bar{\Sigma}_{A2}]$ where the split reflects the distinction between exogenous and endogenous variables. For example, if the second block contains a constant and linear trend, $\bar{\Sigma}_{A2} = \begin{bmatrix} \phi_2 & -0.5\phi_2^2 \\ 0.5\phi_2^2 & -\phi_2^3/3 \end{bmatrix}$), where ϕ_2 is an hyperparameter, while we could set $diag\{\Sigma_{A1i}\} = \phi_0^2 \frac{\phi_1^2}{\ell}$, where ℓ refers to lags, and $\phi_1 = 1$ for lags of the same variable in an equation (Note that, since Σ_{At} is an estimate of the precision, ϕ's should be greater than one). Unless it is required by the problem, set $g_0(A) = 1$. Finally, one could select $\nu \approx 20$ for quarterly data; $\nu \approx 60$ for monthly data and set $\rho = \frac{\nu}{\nu+1}$.

Given the generic structure for the posterior of (A_t, Σ_{t+1}) (a time varying density multiplied by a Normal-Wishart density), we need numerical methods to draw posterior sequences. Any of the approaches described in chapter 9 will do it.

Example 11.6 *An importance sampling approach to draw from this posterior is*

1) *Find the marginal for* A_T. *Integrating* Σ_{T+1} *out of* $\breve{g}(A_T, \Sigma_{t+1}|y)$ *we have* $\breve{g}(A_T|y) = 0.5\sum_t log|(A-\tilde{A}_T)\tilde{\Sigma}_{AT}(A-\tilde{A}_T)' + \frac{\nu}{\rho}\Sigma_T| - 0.5(k+\nu)|(A-\tilde{A}_T)\tilde{\Sigma}_{AT}(A-\tilde{A}_T)' + \frac{\nu}{\rho}\Sigma_T|$.

2) *Find the mode of* $\breve{g}(A_T|y)$ *(call it* A^**) and compute the Hessian at the mode*

3) *Derive the posterior for* Σ_{T+1}^{-1} *(conditional on* A_T, *it is* $\mathbb{W}(\rho((A-\tilde{A}_T)\tilde{\Sigma}_{AT}(A-\tilde{A}_T)' + \nu\tilde{\Sigma}_T), \nu+k)$*).*

4) *Draw* A^l *from a multivariate t-distribution centered at* A^* *and with variance equal to the Hessian at the mode and degrees of freedom* $\nu << T - k(M+1)$. *Draw* $(\Sigma^{-1})^l$ *from the Wishart distribution derived in step 3.*

5) *Calculate importance weights:* $\ln W(A_T^l, \Sigma_{T+1}^l) = constant + \ln(\breve{g}(A^l)) - \ln(\breve{g}^{IS}(A^l))$, *where* $g^{IS}(A^l)$ *is the value of the importance sampling density at* A^l.

6) *Approximate any function* $h(A, \Sigma_t)$ *using* $\bar{h}_L = \frac{\sum_{l=1}^L h(A_T^l, \Sigma_{T+1}^l) W(A_T^l, \Sigma_{T+1}^l)}{\sum_{l=1}^L W(A_T^l, \Sigma_{T+1}^l)}$.

Exercise 11.18 *Describe a MH algorithm to draw posterior sequences for* $(A_T,\ \Sigma_{T+1})$.

Exercise 11.19 *(Cogley) Consider a bivariate model with consumption and income growth of the form* $y_t = \bar{y} + A_t(\ell)y_{t-1} + e_t,\ \alpha_t = vec(A_t(\ell)) = \alpha_{t-1} + v_{1t},\ \Sigma_t = diag\{\sigma_{it}^2\},\ \ln\sigma_t^2 = \ln\sigma_{t-1}^2 + \sigma_{v_2}v_{2t}$, *where* $\bar{y}$ *is a constant. In a constant coefficient version of the model the trend growth rate of the two variables is* $(I - A(\ell))^{-1}\bar{y}$. *Using a Gibbs sampler, describe how to construct a time varying estimate of the trend growth rate,* $(I - A_t(\ell))^{-1}\bar{y}$.

We conclude this section applying Bayesian methods to the estimation of the parameters of a GARCH model.

Example 11.7 *Consider the model* $y_t = x_t'A + \sigma_t e_t, \;\; e_t \sim \mathbb{N}(0,1)$ *and* $\sigma_t^2 = \rho_0 + \rho_1\sigma_{t-1}^2 + \rho_2 e_{t-1}^2$. *Assume* $A \sim \mathbb{N}(\bar{A}, \bar{\sigma}_A^2)$, $\rho_0 \sim \mathbb{N}(\bar{\rho}_0, \bar{\sigma}_{\rho_0}^2)$ *and that the prior for* ρ_1, ρ_2 *is uniform over [0,1] and restricted so that* $\rho_1 + \rho_2 \leq 1$. *The posterior kernel can be easily constructed from these densities. Let* $\alpha = (A, \rho_i, i = 0,1,2)$*; let the mode of the posterior be* α^**.and let* $\breve{t}(.)$ *be the kernel of a t-distribution with location* α^**, scale proportional to the Hessian at the mode and* $\bar{\nu}$ *degrees of freedom. Posterior draws for the parameters can be obtained using an independence Metropolis algorithm, i.e. generate* $\alpha^\ddagger$ *from* $\breve{t}(.)$ *and accept the draw with probability equal to* $\min[\frac{\breve{g}(\alpha^\ddagger|y_t)/\breve{t}(\alpha^\ddagger)}{\breve{g}(\alpha^{l-1}|y_t)/\breve{t}(\alpha^{l-1})}, 1]$. *A t-distribution is appropriate in this case because* $\frac{\breve{g}(\alpha|y_t)}{\breve{t}(\alpha)}$ *is typically bounded from above.*

11.3 Markov switching models

Markov switching models are extensively used in macroeconomics, in particular, when important relationships are suspected to be functions of an unobservable variable (e.g. the state of business cycle). Hamilton (1994) provides a classical nonlinear filtering method to obtain estimates of the parameters and of the unobservable state. Here we consider a Bayesian approach to the problem. As with factor and stochastic volatility models, the unobservable state is treated as "missing" data and sampled together with other parameters using the Gibbs sampler.

To set up ideas we start from a static model where the slope varies with the state:

$$y_t \;\; = \;\; x_{1t}A_1 + x_{2t}A_2(\varkappa_t - 1) + e_t \quad\quad e_t \sim \mathbb{N}(0, \sigma_e^2) \tag{11.18}$$

Here $\varkappa_t$ is a two-state Markov switching indicator. We take $\varkappa_t = 1$ to be the normal state so that $y_t = x_{1t}A_1 + v_t$. In the extraordinary state, $\varkappa_t = 0$ and $y_t = x_{1t}A_1 - x_{2t}A_2 + e_t$.

We let $p_1 = P(\varkappa_t = 1|\varkappa_{t-1} = 1)$; $p_2 = p(\varkappa_t = 0|\varkappa_{t-1} = 0)$, both of which are unknown; also we let $y^{t-1} = (y_1, \ldots, y_{t-1}, x_{i1}, \ldots, x_{it-1}, \; i = 1,2)$, $\varkappa^t = (\varkappa_1, \ldots, \varkappa_t)$, $\alpha = (A_1, A_2, \sigma_e^2, \varkappa^t, p_1, p_2)$. We want to obtain the posterior distribution for α. We assume $g(\alpha) = g(A_1, A_2, \sigma_e^2)g(\varkappa^t|p_1, p_2)g(p_1, p_2)$. We let $g(p_1, p_2) = p_1^{\bar{d}_{11}}(1-p_2)^{\bar{d}_{12}}p_2^{\bar{d}_{22}}(1-p_2)^{\bar{d}_{21}}$ where $\bar{d}_{ij}$ are the a-priori proportions of the (i,j) elements in the sample. As usual, we assume $g(A_1, A_2, \sigma_e^{-2}) \propto \mathbb{N}(\bar{A}_1, \bar{\Sigma}_1) \times \mathbb{N}(\bar{A}_2, \bar{\Sigma}_2) \times \mathbb{G}(a_1, a_2)$.

The posterior kernel is $\breve{g}(\alpha|y) = \sum_{t=1}^T f(y_t|\alpha, y^{t-1})g(\alpha)$ where $f(y_t|\alpha, y^{t-1}) \sim N(Ax_t, \sigma_e^2)$, $x_t = (x_{1t}, x_{2t})$ and $A = (A_1, A_2)$. To sample from this kernel we need starting values for α and $\varkappa_t$ and the following algorithm:

Algorithm 11.2

1) *Sample* p_i *from* $g(p_1, p_2|y) = p_1^{\bar{d}_{11}+d_{11}}(1-p_1)^{\bar{d}_{12}+d_{12}}p_2^{\bar{d}_{22}+d_{22}}(1-p_2)^{\bar{d}_{21}+d_{21}}$, *where* d_{ij} *is the number of actual shifts between state* i *and state* j.

2) *Sample* A_i *from* $\breve{g}(A_i|\sigma_e^2, \varkappa^T, y)$. *This kernel is normal with variance* $\tilde{\Sigma}_A = \sum_t \frac{x_t' x_t}{\sigma^2} + \bar{\Sigma}^{-1})^{-1}$ *and mean* $\tilde{A} = \tilde{\Sigma}_A(\sum_t \frac{x_t y_t}{\sigma^2} + \bar{\Sigma}_A^{-1}\bar{A})$, *where* $\bar{A} = (\bar{A}_1, \bar{A}_2)$ *and* $\bar{\Sigma} = \text{diag}(\bar{\Sigma}_1, \bar{\Sigma}_2)$.

3) *Sample* σ_e^{-2} *from* $\breve{g}(\sigma_e^{-2}|\varkappa^T, y, A)$. *This is the kernel of a Gamma with parameters equal to* $a_1 + (T-1)/2$ *and* $a_2 + 0.5\sum_t (y_t - A_1 x_{1t} + A_2 x_{2t}(\varkappa_t - 1))^2$.

4) *Sample* $\varkappa^T$ *from* $\breve{g}(\varkappa^T|y, A, \sigma_e^2, p_i)$. *As usual we do this in two steps. Given* $g(\varkappa_0)$ *we run forward into the sample using* $g(\varkappa_t|A, \sigma_e^2, y^t, p_i) \propto f(y_t|y^{t-1}, A, \sigma_e^2, \varkappa_t) g(\varkappa_t|A, \sigma_e^2, y^{t-1}, p_i)$ *where* $f(y_t|y^{t-1}, A, \sigma_e^2, \varkappa_t) \sim N(Ax_t, \sigma_e^2)$ *and* $g(\varkappa_t|A, \sigma_e^2, y^{t-1}, p_i) = \sum_{\varkappa_{t-1}=0}^{1} g(\varkappa_{t-1}|A, \sigma_e^2, y^{t-1}, p_i) \times P(\varkappa_t = i|\varkappa_{t-1} = j)$. *Then, starting from* $\varkappa_T$, *we run backwards to smooth estimates, i.e. given* $g(\varkappa_T|y^T, A, \sigma_e^2, p_i)$, *we compute* $g(\varkappa_\tau|\varkappa_{\tau+1}, y^\tau, A, \sigma_e^2, p_i) \propto g(\varkappa_\tau|A, \sigma_e^2, y^\tau, p_i) P(\varkappa_\tau = i|\varkappa_{\tau+1} = j)^{-1}$, $\tau = T-1, T-2, \ldots$. *Note that we have used the Markov properties of* $\varkappa_t$ *to split the forward and backward problems of drawing* T *joint values into the problem of drawing* T *conditional values.*

It is immediate to recognize that step 4) of the algorithm 11.2 is the same as the one we have used to extract the unobservable state in state space models and the amount of computation involved is similar. In fact, the first part is similar to drawing the AR parameters in a factor model and the second to the estimation of the factor at each stage of the simulation. This is not surprising: a two-state Markov chain model can always be written as a first order AR process with AR coefficient equal to $p_2 + p_1 - 1$. The difference, as already mentioned, is that the AR process has binary innovations.

Exercise 11.20 *Suppose that the prior for* p_i *is non-informative. Show the form of the conditional posterior of* $(A_1, A_2, \sigma_e^{-2})$. *Alter algorithm 11.2 to take into account this change.*

Example 11.8 *We use equation (11.18) to study fluctuations in EU industrial production. To construct a EU measure we aggregate IP data for Germany, France and Italy using GDP weights and let* y_t *be the yearly changes in industrial production. Data runs from 1974:1 to 2001:4. The posterior means are* $\tilde{A}_2 = 0.46$ *and* $\tilde{A}_1 = 0.96$ *and the standard deviations equals 0.09 and 0.09, respectively. Hence, the annual growth rate in expansions is about two percentage points higher and the difference is statistically significant. Estimates of the probability of being in the extraordinary state (a "recession") are in figure 11.1: the algorithm picks up the standard recessions (1975, 1980, 1982, 1993) and indicates the presence of a new contractionary phase starting in 2001:1.*

11.3.1 A more complicated structure

The model we consider here is:

$$A^y(\ell)(y_t - \bar{y}(\varkappa_t, x_t)) = \sigma(\varkappa_t) e_t \tag{11.19}$$

where $A^y(\ell)$ is a polynomial in the lag operator, $\bar{y}(\varkappa_t, x_t)$ is the mean of y_t, which depends on observable regressors x_t and on the unobservable state $\varkappa_t$, var(e_t)=1, $\sigma^2(\varkappa_t)$ also depends

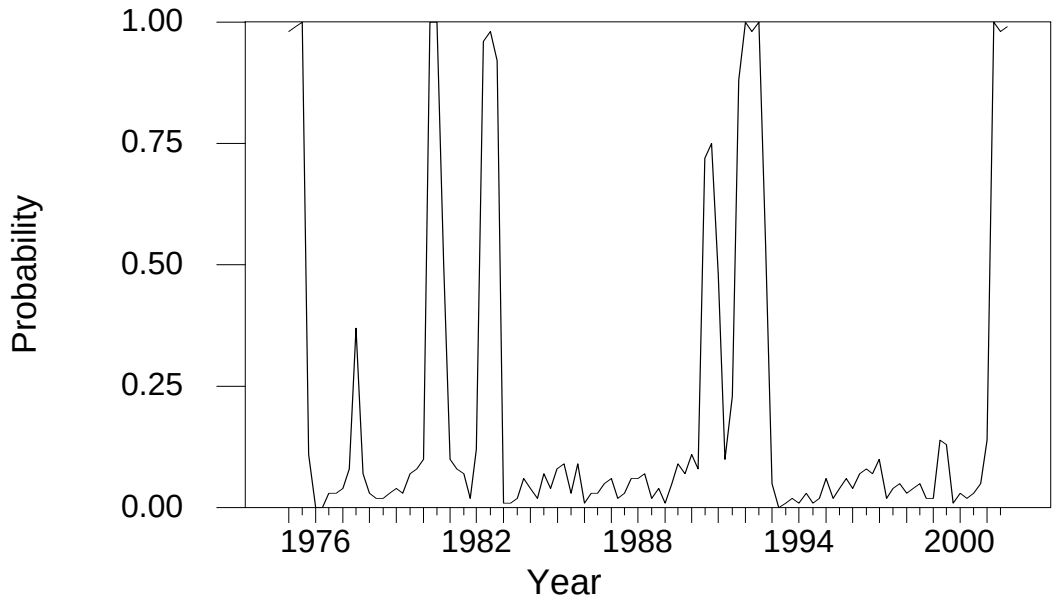


Figure 11.1: Recession probabilities.

on the unobservable state, and $\varkappa_t$ is a two-state Markov chain with transition matrix P. We specify $\bar{y}(\varkappa_t, x_t) = x_t' A_0 + A_1 \varkappa_t$, $\sigma(\varkappa_t) = \sigma^2 + A_2 \varkappa_t$ and assume $A_2 > 0,\ A_1 > 0$ since, for $A_2 = A_1 = 0$, the two states are not identified. We restrict the roots of $A^y(\ell)$ to be less than one.

Let $y^t = (y_1, \dots y_t)$, $\varkappa^t = (\varkappa_1, \dots, \varkappa_t)$; let $\mathbb{A}$ be the companion matrix of the AR polynomial $A^y(\ell)$, and $\mathbb{A}_1$ its first m rows. Define $\kappa = \frac{A_2}{\sigma^2}$ and let $\alpha = (A_0, A_1, \mathbb{A}_1, \sigma^2, \kappa, p_{ij})$. The likelihood function is $f(y^t|\varkappa^t, \alpha) = f(y^q|\varkappa^q, \alpha) \prod_{\tau=q+1}^{t} f(y_\tau|y^{\tau-1}, \varkappa^{t-1}, \alpha)$, where the first term is the density of the first q observations and the second the one-step ahead conditional density of y_j.

The density of the first q observations (see derivation in the factor model case) is normal with mean $x^q A_0 + \varkappa^q A_1$ and variance $\sigma^2 \Omega_q$, where $\Omega_q = W_q \Sigma_q W_q$, $\Sigma_q = \mathbb{A}\Sigma_q \mathbb{A}' + (1, 0, 0, \dots, 0)'(1, 0, 0, \dots, 0)$, $W_q = \text{diag}\{(1 + \kappa\varkappa_j)^{0,5}\ j = 1, \dots, q\}$. Using the prediction error decomposition we have that $f(y_\tau|y^{\tau-1}, \varkappa^{\tau-1}, \alpha)$ is proportional to $\exp\{-0.5\sigma^{-2}(\varkappa_\tau)^(y_\tau - y_{\tau|\tau-1})^2\}$ where $y_{\tau|\tau-1} = (1 - A^y(\ell))y_t + A^y(\ell)(x_\tau' A_0 + A_1 \varkappa_\tau)$. Therefore, y_t is conditionally normal with mean $y_{t|t-1}$ and variance $\sigma^2(\varkappa_t)$. Finally, the joint density of $(y^t, \varkappa^t)$ is equal to $f(y^t|\varkappa^t, \alpha)\ \prod_{\tau=2}^{t} f(\varkappa_\tau|\varkappa_{\tau-1}) f(\varkappa_1)$ and the likelihood of the data is $\int f(y^t, \varkappa^t|\alpha) d\varkappa^t$. In chapter 3 we have produced estimates of $(\alpha, \varkappa^t)$ using a two-step approach where in the first step α_{ML} is obtained maximizing the likelihood function. In the second step, inference

about $\varkappa^t$ is obtained conditional on α_{ML}. That is,

$$\begin{aligned} f(\varkappa_t,\ldots,\varkappa_{t-\tau+1}|y^t,\alpha_{ML}) &= \sum_{\varkappa_{t-\tau}=0}^{1} f(\varkappa_t,\ldots,\varkappa_{t-\tau}|y^{t-1},\alpha_{ML}) \\ &\propto f(\varkappa_t|\varkappa_{t-1})f(\varkappa_{t-1},\ldots\varkappa_{t-\tau}|y^{t-1},\alpha_{ML})f(y_t|y^{t-1},\varkappa^t,\alpha_{ML}) \end{aligned} \tag{11.20}$$

where the factor of proportionality is given by $f(y_t|y^{t-1},\alpha_{ML})=\sum_{\varkappa_t},\ldots\sum_{\varkappa_{t-\tau}} f(y_t,\varkappa_t,\ldots,\varkappa_{t-\tau}|y^{t-1},\alpha_{ML})$. Since the log likelihood of the sample is $\log f(y_{q+1},\ldots y_t|y^q,\alpha)=\sum_\tau \log f(y_\tau|y^{\tau-1},\alpha)$, once α_{ML} is obtained, transition probabilities can be computed using $f(\varkappa_t|y^t,\alpha_{ML})=\int\ldots\int f(\varkappa_t,\ldots,\varkappa_{t-\tau+1}|y^t,\alpha_{ML})d\varkappa_{t-1}\ldots d\varkappa_{t-\tau+1}$. Note that uncertainty in α_{ML} is not incorporated in the calculations.

To construct the conditional posteriors of the parameters and of the unobservable state, assume that $g(A_0,A_1,\sigma^{-2})\propto \mathbb{N}(\bar{A}_0,\bar{\Sigma}_{A_0})\times\mathbb{N}(\bar{A}_1,\bar{\Sigma}_{A_1})\mathcal{I}_{(A_1>0)}\times\mathbb{G}(a_1^\sigma,a_2^\sigma)$ where $\mathcal{I}_{(A_1>0)}$ is an indicator function. Further assume that $g((1+\kappa)^{-1})\sim\mathbb{G}(a_1^\kappa,\ a_2^\kappa)\mathcal{I}_{(\kappa>0)}$, that $g(\mathbb{A}_1)\sim\mathbb{N}(\bar{\mathbb{A}}_1,\bar{\Sigma}_{A_1})\mathcal{I}_{[-1,1]}$ where $\mathcal{I}_{[-1,1]}$ is an indicator for stationarity. Finally, we let $p_{12}=1-p_{11}=1-p_1$ and $p_{21}=1-p_{22}=1-p_2$ and $g(p_i)\propto Beta(\bar{d}_{i1},\bar{d}_{i2}),\ i=1,2$, and assume that all hyperparameters are known.

Exercise 11.21 *Let $\alpha_{-\psi}$ be the vector of parameters α except for ψ and let $A=(A_0,A_1)$.*
i) Repeating the same steps outlined in the previous sections, assuming that the first q observations come from the low state, show that the conditional posteriors for the parameters and the unobserved state are

$$\begin{aligned} g(A|y^t,\varkappa^t,\alpha_{-A}) &\sim \mathbb{N}(\tilde{A},\tilde{\Sigma}_A)\mathcal{I}_{A_1>0} \\ g(\sigma^{-2}|y^t,\varkappa^t,\alpha_{-\sigma_e^2}) &\sim \mathbb{G}(a_1^\sigma+T,a_2^\sigma+(\Sigma_q^{-0.5}y-\Sigma_q^{-0.5}xA_0+\Sigma_q^{-0.5}\varkappa A_1)^2) \\ g((1+\kappa)^{-1}|y^t,\varkappa^t,\alpha_{-\kappa}) &\sim \mathbb{G}(a_1^\kappa+T_1,a_2^\kappa+rss)\mathcal{I}_{(\kappa>0)} \\ g(\mathbb{A}_1|y^t,\varkappa^t,\alpha_{-\mathbb{A}_1}) &\sim \mathbb{N}(\tilde{\mathbb{A}}_1,\tilde{\Sigma}_{\mathbb{A}_1})\mathcal{I}_{[-1,1]}|\Omega_q|^{-0.5}\exp\{-0.5\sigma^{-2}(y^q-x^qA)'\Omega_q^{-1}(y^q-x^qA)\} \\ g(p_i|y^t,\varkappa^t,\alpha_{-p}) &\sim \text{Beta}(\bar{d}_{i1}+d_{i1},\bar{d}_{i2}+d_{i2}) \quad i=1,2 \\ g(\varkappa_t|y^t,\alpha_{-\varkappa_{-t}}) &\propto f(\varkappa_t|\varkappa_{t-1})f(\varkappa_{t+1}|\varkappa_t)\prod_\tau f(y_\tau|y^{\tau-1},\varkappa^\tau) \end{aligned} \tag{11.21}$$

where T_1 is the number of elements in T for which $\varkappa_t=1$, d_{ij} is the number of transitions from state i to state j and $rss=\sum_{t=1}^{T_1}[\frac{(1+\kappa\varkappa_t)^{0.5}y_t-(1+\kappa\varkappa_t)^{0.5}x_tA_0-(1+\kappa\varkappa_t)^{0.5}\varkappa_tA_1}{\sigma}]^2$.
ii) Show the exact form of $\tilde{\mathbb{A}}_1,\tilde{\Sigma}_{A_1},\tilde{A},\tilde{\Sigma}_A$.
iii) Describe how to produce draws for $\mathbb{A}_1$ and A restricted to the correct domain.

Recently, Sims (2001) Sims and Zha (2004) have used a similar specification to estimate a Markov switching VAR model where the switch may occur in the lagged dynamics, in the contemporaneous effects, or in both. To illustrate their approach consider the equation

$$A_1(\ell)i_t=\bar{i}(\varkappa_t)+b(\varkappa_t)A_2(\ell)\pi_t+\sigma(\varkappa_t)e_t \tag{11.22}$$

where $e_t \sim \mathbb{N}(0,1)$, i_t is the nominal interest rate, π_t is the commodity price inflation and $\varkappa_t$ has three states with transition $P = \begin{bmatrix} p_1 & 1-p_1 & 0 \\ 0.5*(1-p_2) & p_2 & 0.5*(1-p_2) \\ 0 & 1-p_3 & p_3 \end{bmatrix}$. The model (11.22) imposes restrictions on the data: the dynamics of interest rates do not depend on the state; the form of the lag distribution on π_t is the same across states, except for a scale factor $b(\varkappa)$; there is no possibility to jump from state 1 to state 3 (and viceversa) without passing through state 2; finally, the nine elements of P depend only on three parameters.

Let $\alpha = vec(A_1(\ell)), vec(A_2(\ell)), \bar{i}(\varkappa_t), b(\varkappa_t), \sigma(\varkappa_t), p_1, p_2, p_3)$. The marginal likelihood of the data, conditional on the parameters (but integrating out the unobservable state) can be computed numerically and recursively. Let $\mathcal{F}_t$ be the information set at t

Exercise 11.22 *Show that $f(i_t, \varkappa_t|\mathcal{F}_{t-1})$ is a mixture of continuous and discrete densities. Show the form of $f(i_t|\mathcal{F}_{t-1})$, the marginal of the data, and $f(\varkappa_t|\mathcal{F}_t)$, the updating density.*

Once $f(\varkappa_t|\mathcal{F}_t)$ is obtained we can compute $f(\varkappa_{t+1}|\mathcal{F}_t) = \begin{bmatrix} f(\varkappa_t = 1|\mathcal{F}_t) \\ f(\varkappa_t = 2|\mathcal{F}_t) \\ f(\varkappa_t = 3|\mathcal{F}_t) \end{bmatrix}' P$ and from there we can calculate $f(i_{t+1}, \varkappa_{t+1}|i_t, \pi_t, \ldots)$ which makes the recursion complete. Given a flat prior on α, the posterior will be proportional to $f(\alpha|i_t, \pi_t)$ and posterior estimates of the parameters and of the states can be immediately obtained.

Exercise 11.23 *Provide formulas to obtain smoothed estimates of $\varkappa_t$.*

More complicated VAR specifications are possible. For example, let $y_t\mathcal{A}_0(\varkappa_t) = x_t'\mathcal{A}_+(\varkappa_t) + e_t$, where x_t includes all lags of y_t and $e_t \sim \mathbb{N}(0, I)$. Then, as we have done in chapter 10, assume $\mathcal{A}_+(\varkappa_t) = \mathcal{A}(\varkappa_t) + [I, 0]'\mathcal{A}_0(\varkappa_t)$. Given this specification there are two possibilities: (i)$\mathcal{A}_0(\varkappa_t) = \bar{A}_0\Lambda(\varkappa_t)$ and $\mathcal{A}(\varkappa_t) = \bar{\mathcal{A}}\Lambda(\varkappa_t)$ or (ii) $\mathcal{A}_0(\varkappa_t)$ free and $\mathcal{A}(\varkappa_t) = \bar{\mathcal{A}}$. In the first specification changes in the contemporaneous and lagged coefficients are proportional; in the second the state affects the contemporaneous relationship but not lagged ones.

Equation (11.22) is an equation of a bivariate VAR. Hence, so long as we are able to keep the posterior of the system in a SUR format (as we have done in chapter 10), the above ideas can be applied to each of the VAR equations.

11.3.2 A General Markov switching specification

Finally, we consider a general Markov switching specification which embeds as special case the two previous ones. So far we have allowed the mean and the variance of the process for y_t to be switching with the state but we have forced the dynamics to be independent of the state, apart from a scale effect. This is a strong restriction: it is conceivable, e.g., that the autocovariance functions is different in expansions and in recessions.

The general two-state Markov switching model we consider here is:

$$\begin{aligned} y_t &= x_t'A_{01} + Y_t'A_{02} + e_{0t} \qquad \text{if } \varkappa_t = 0 \\ &= x_t'A_{02} + Y_t'A_{12} + e_{1t} \qquad \text{if } \varkappa_t = 1 \end{aligned} \tag{11.23}$$

where x_t is a $1 \times q_2$ vector of exogenous variables for each t, $Y_t' = (y_{t-1}, \ldots y_{t-q_1})$ is a vector of lagged dependent variables and e_{jt}, $j = 0, 1$, are iid random variables, normally distributed with mean zero and variance σ_j^2. Once again the transition probability for $\varkappa_t$ has diagonal elements p_i. In principle, some of the elements of A_{ji} may be equal to zero for some i, so the model may have different dynamics in different states.

Without further restrictions the two states are not identified. To achieve identification, choose the first state as a "recession", so that $A_{02} < A_{12}$ is imposed. We let α_c be the parameters which are common across states, α_i the parameters which are unique to the state and α_{ir} the parameters which are restricted to achieve identification. Then (11.23) can be written as

$$\begin{aligned} y_t &= X_{ct}'\alpha_c + X_{0t}'\alpha_0 + X_{rt}'\alpha_{0r} + e_{0t} \qquad \text{if } \varkappa_t = 0 \\ &= X_{ct}'\alpha_c + X_{1t}'\alpha_1 + X_{rt}'\alpha_{1r} + e_{1t} \qquad \text{if } \varkappa_t = 1 \end{aligned} \tag{11.24}$$

where $(X_{ct}', X_{it}', X_{rt}') = (x_t', Y_t')$ and $(\alpha_c, \alpha_i, \alpha_{ir}) = (A_{01}, A_{02}, A_{11}, A_{12})$.

To construct conditional posteriors for the parameters we assume conjugate priors: $\alpha_c \sim \mathbb{N}(\bar{\alpha}_c, \bar{\Sigma}_c)$; $\alpha_i \sim \mathbb{N}(\bar{\alpha}_i, \bar{\Sigma}_i)$; $\alpha_{ir} \sim \mathbb{N}(\bar{\alpha}_r, \bar{\Sigma}_r)\mathcal{I}_{(rest)}$; $\bar{s}_i^2\sigma_i^{-2} \sim \chi^2(\bar{\nu}_i)$; $p_i \sim \text{Beta}(\bar{d}_{1i}, \bar{d}_{2i})$ $i = 1, 2$ where $\mathcal{I}_{(rest)}$ is a function indicating whether the identification restrictions are satisfied. As usual we assume that the hyperparameters $(\bar{\alpha}_c, \bar{\Sigma}_c, \bar{\alpha}_i, \bar{\Sigma}_i, \bar{\alpha}_r, \bar{\Sigma}_r, \bar{\nu}_i, \bar{s}_i^2, \bar{d}_{ji})$ are known or can be estimated from the data. We take the first $max[q_1, q_2]$ observations as given in constructing the posterior distribution of the parameters and of the latent variable.

Given these priors, it is straightforward to compute conditional posteriors. For example, the conditional posterior of α_c is normal with mean $\tilde{\alpha}_c = \tilde{\Sigma}_c(\sum_{t=min[q_1,q_2]}^{T} \frac{X_{ct}y_{c,t}'}{\sigma_t^2} + \bar{\Sigma}_c^{-1}\bar{\alpha}_c)$, where $y_{c,t} = y_t - X_{it}\alpha_i - X_{rt}\alpha_{ir}$, and variance $\tilde{\Sigma}_c = (\sum_{t=min[q_1,q_2]}^{T} \frac{X_{ct}X_{ct}'}{\sigma_t^2} + \bar{\Sigma}_c^{-1})^{-1}$.

Exercise 11.24 *Let T_i is the number of observations in state i. Show that:*
i) the conditional posterior of α_i is $\mathbb{N}(\tilde{\alpha}_i, \tilde{\Sigma}_i)$ where $\alpha_i = \tilde{\Sigma}_i(\sum_{t=1}^{T_i} \frac{X_{it}y_{i,t}'}{\sigma_t^2} + \bar{\Sigma}_i^{-1}\bar{\alpha}_i)$, $y_{i,t} = y_t - X_{ct}\alpha_c - X_{rt}\alpha_{ir}$, $\tilde{\Sigma}_i = (\sum_{t=1}^{T_i} \frac{X_{it}X_{it}'}{\sigma_t^2} + \bar{\Sigma}_i^{-1})^{-1}$.
ii) the conditional posterior of α_r is $\mathbb{N}(\tilde{\alpha}_r, \tilde{\Sigma}_r)$. Show the form of $\tilde{\alpha}_r, \tilde{\Sigma}_r$.
iii)The conditional distribution of σ_i^{-2} is such that $\frac{(\bar{s}_i^2 + rss_i^2)}{\sigma_i^2} \sim \chi^2(\nu_i + T_i - \max[q_1, q_2])$. Write down the expression for rss_i^2.
iv) The conditional posterior for p_i is Beta$(\bar{d}_{1i} + d_{1i}, \bar{d}_{2i} + d_{2i})$.

Finally, the conditional posterior for the latent variable $\varkappa_t$ can be computed as usual. Given the Markov properties of the model, we restrict attention to the subsequence $\varkappa_{t,\tau} = (\varkappa_t, \ldots, \varkappa_{t+\tau-1})$. Define $\varkappa_{t(-\tau)}$ as the sequence $\varkappa_t$ with the $\tau - th$ subsequence removed. Then $g(\varkappa_{t,\tau}|y, \varkappa_{t(-\tau)}) \propto f(y|\varkappa_t, \alpha, \sigma^2)g(\varkappa_{t,\tau}|\varkappa_{t(-\tau)}, p_i)$, which is a discrete distribution with 2^τ outcomes. Using the Markov property, $g(\varkappa_{t,\tau}|\varkappa_{t(-\tau)}, p_i) = g(\varkappa_{t,\tau}|\varkappa_{t-1}, \varkappa_{t+\tau}, p_i)$ while $f(y^T|\varkappa_t, \alpha) \propto \prod_{j=t}^{t+\tau-1} \frac{1}{\sigma_j} \exp\{-0.5\frac{e_j^2}{\sigma_j^2}\}$. Since the $\varkappa_t$ are correlated, it is a good idea to choose $\tau > 1$, but the above can be easily applied to the case $\tau = 1$.

Exercise 11.25 *Show the components of the conditional posterior for $\varkappa_t$ when $\tau = 1$.*

In all Markov switching specifications, it is important to wisely select the initial conditions. One way to do so is to assign all the observations in the training sample to one state, obtain initial estimates for the parameters and arbitrarily set the parameters of the other state to be equal to the estimates plus or minus a small number (say, 0.1). Alternatively, one can split the points arbitrarily but equally across the two states.

Exercise 11.26 *Suppose $\Delta y_t = \alpha_0 + \alpha_1 \Delta y_{t-1} + e_t$, where $e_t \sim \mathbb{N}(0, \sigma_e^2)$ if $\varkappa_t = 0$ and $\Delta y_t = (\alpha_0 + A_0) + (\alpha_1 + A_1)\Delta y_{t-1} + e_t$ where $e_t \sim \mathbb{N}(0, (1 + A_2)\sigma_e^2)$ if $\varkappa_t = 1$. Using quarterly EU GDP data, construct posterior estimates for A_0, A_1, A_2. Separately test if there is evidence of switching in the intercept, in the dynamics or in the variance of the process.*

11.4 Bayesian DSGE Models

The use of Bayesian methods to estimate and evaluate Dynamic Stochastic General Equilibrium (DSGE) models does not present new theoretical aspects. We have repeatedly mentioned that DSGE models are false in at least two senses:

- They provide only an approximate representation to the DGP. Since the vector of structural parameters is of low dimension, strong restrictions are implied both in the short and in the long run.
- The number of driving forces is smaller than the number of endogenous variables so that the covariance matrix of a vector of variables generated by the model is singular.

These basic features make the estimation and testing of DSGE models with GMM or ML tricky. In Chapter 4 we have described a minimalist approach, which uses qualitative restrictions to identify shocks in the data, and can be employed to examine the match between the theory and the data, when the model is false in the two above senses.

Bayesian methods are also well suited to deal with false models. Posterior inference, in fact, does not hinge on the model being the correct DGP and it is feasible even when the covariance matrix of the vector of endogenous variables is singular. Bayesian methods have another advantage over alternatives, which make them appealing to macroeconomists. The posterior distribution of the statistics of interest incorporates prior uncertainty, both about the parameters and the model specification.

Since log-linearized DSGE models are state space models with nonlinear restrictions on the mapping between reduced form and structural parameters, posterior estimates of the structural parameters can be obtained, for appropriately designed prior distributions, using the posterior simulators described in chapter 9. Given the non-linearity of the mapping, it is difficult to build the conditional distributions used in the Gibbs sampler. For that reason, Metropolis or MH algorithms are generally employed. Numerical methods can also be used

to compute predictive densities and Bayes factors; to obtain any posterior function of the structural parameters (in particular, impulse responses, variance decompositions, the ACF, turning point predictions, and forecasts, etc.) and to examine the sensitivity of the results to variations in the prior specification. Once the posterior distribution of the structural parameters is obtained, it becomes trivial to conduct any inferential exercise a researcher is interested in.

To estimate the posterior for the structural parameters and for the statistics of interest, and to evaluate the quality of a DSGE model the following steps are typically used.

Algorithm 11.3

1) *Construct a log-linear representation of the DSGE economy and transform it into a state space model. Add measurement errors if evaluation/estimation is done on a vector of variables which is larger than the one of the states.*

2) *Specify prior distributions for the structural parameters θ.*

3) *Compute the predictive density numerically, using draws from the prior distribution and the Kalman filter. Compute the predictive density for any alternative or reference model. Calculate Bayes factors or other measures of (relative) forecasting fit.*

4) *Draw sequences from the joint posterior of the parameters using Metropolis or MH algorithms. Check convergence.*

5) *Construct statistics of interest using the draws in 4) (after an initial set has been discarded). Use loss-based measures to evaluate the discrepancy between the theory and the data.*

6) *Examine sensitivity of the results to the choice of priors.*

Step 1) is unnecessary. We will see later on what to do if a nonlinear specification is used. Adding measurement errors helps computationally to reduce the singularity of the covariance matrix of the endogenous variables but it is not needed for the approach to work.

In step 2) prior distributions are centered around standard calibrated values of the parameters while standard errors generally reflect subjective prior uncertainty faced by an investigator. One could also specify standard errors so as to "cover" the range of existing estimates, as we have done in chapter 7. In some applications, it may be convenient to select diffuse priors over a fixed range to avoid to put too much structure on the data. In general, the form of the prior reflects computation convenience. Conjugate priors are typically preferred. For autoregressive parameters or parameters which must lie in an interval, truncated Normal or Beta distributions are chosen.

Step 3) requires drawing parameters from the prior specified in 2), calculating the sequence of prediction errors for each draw and averaging over draws. That is, we numerically estimate the predictive density using e.g. the modified harmonic mean suggested by Gelfand and Dey (1994), $[\frac{1}{L}\sum_l(\frac{g^{IS}(\theta^*)}{f(y|\theta^*)g(\theta^*)}$ for some high probability θ^* where g^{IS} is a density with

tail thinner than $f(y|\theta)g(\theta)$, or directly using the Bayes theorem as suggested in Chib (1995). Similar calculations can be undertaken for any alternative model and Bayes factors can then be numerically computed. When the dimensionality of the parameters space is large, it may be convenient to use Laplace approximations to reduce the computation burden. The competitors could be structural models, which nest the one under consideration (e.g. a model with flexible prices can be obtained with a restriction on one parameter of a model with sticky prices); non-nested structural specifications (e.g. a model with sticky wages) or more densely parametrized reduced form models (e.g. VAR or a BVAR).

Steps 4)-5) require choosing an updating rule, and a transition function $\mathsf{P}(\theta^{\ddagger}, \theta^{l-1})$ satisfying the regularity conditions described in chapter 9; estimating joint and marginal distributions using kernel methods and the draws from the posterior; and setting up a loss function reflecting the costs faced by an investigator in selecting a model. In particular, the following steps are needed:

Algorithm 11.4

1) *Given a* θ^0*, draw* $\theta^{\ddagger}$ *from* $\mathsf{P}(\theta^{\ddagger}, \theta^0)$.

2) *Use filtering techniques to compute the prediction error decomposition of the likelihood.*

3) *Evaluate the posterior kernel at* $\theta^{\ddagger}$ *and at* θ^0 *i.e. calculate* $\breve{g}(\theta^{\ddagger}) = f(y|\theta^{\ddagger})g(\theta^{\ddagger})$ *and* $\breve{g}(\theta^0) = f(y|\theta^0)g(\theta^0)$.

4) *Draw* $\mathsf{U} \sim \mathbb{U}(0,1)$. *Set* $\theta^1 = \theta^{\ddagger}$ *if* $\mathsf{U} < min[\frac{\breve{g}(\theta^{\ddagger})}{\breve{g}(\theta^0)} \frac{\mathsf{P}(\theta^0, \theta^{\ddagger})}{\mathsf{P}(\theta^{\ddagger}, \theta^0)}, 1]$, *otherwise set* $\theta^1 = \theta^0$.

5) *Iterate on steps 1)-4)* $\bar{L} + JL$ *times. Discard the first* $\bar{L}$ *draws, keep one draw every* L *for inference. Alternatively, repeat iteration on steps 1)-4) using different* θ^0, $L+1$ *times and keep the last draw from each run.*

6) *Estimate marginal/ joint posteriors using kernel methods. Compute location estimates and credible sets, compare them with those obtained from the prior.*

7) *Compute any economically interesting function of the posterior of* θ. *Set up a loss function. Compare the economic quality of a model using the corresponding risk function.*

In step 6, to check the robustness of the results to the choice of prior, one can reweigh the posterior draws using the techniques described in chapter 9, section 5.

Next, we present a few examples, highlighting the steps needed to use Bayesian methods for inference in DSGE models.

Example 11.9 *The first example is very simple. We simulate data from a RBC model where the solution is contaminated by measurement errors. Armed with reasonable prior specifications for the structural parameters and a MH algorithm, we then ask where the posterior distribution of some crucial parameters lies relative to the "true" parameters we*

used in the simulations when samples typical in macroeconomic data are available. We also compare true and estimated moments to give an economic sense to the fit we obtain.

The solution to a RBC model driven by iid technological disturbances when capital depreciates instantaneously, leisure does not enter the utility function and the latter is logarithmic in consumption was obtained in chapter 2, which we repeat for convenience:

$$\begin{aligned} K_{t+1} &= (1-\eta)\beta K_t^{1-\eta}\zeta_t + v_{1t} && (11.25)\\ GDP_t &= K_t^{1-\eta}\zeta_t + v_{2t} && (11.26)\\ c_t &= \eta\beta GDP_t + v_{3t} && (11.27)\\ r_t &= (1-\eta)\frac{GDP_t}{K_t} + v_{4t} && (11.28) \end{aligned}$$

We have added four measurement errors $v_{jt},\ j = 1,2,3,4$ to the equations to reduce the singularity of the system and to mimic the typical situation an investigator is likely to face. Here β is the discount factor, $1-\eta$ the share of capital in production, and σ_ζ is the standard error of the logarithm of the technological disturbance. We simulate 1000 data points using $k_0 = 100.0; (1-\eta) = 0.36; \beta = 0.99,\ \ln\zeta_{1t} \sim \mathbb{N}(0, \sigma_\zeta^2 = 0.1),\ v_{1t} \sim \mathbb{N}(0, 0.06); v_{2t} \sim \mathbb{N}(0, 0.02); v_{3t} \sim \mathbb{N}(0, 0.08), v_{4t} \sim \mathbb{U}(0, 0.1)$ and keep only the last 160 data points to reduce the dependence on the initial conditions and to match existing sample sizes.

We treat σ_ζ^2 as fixed and, for illustrative purposes, focus attention on the two economic parameters. We assume that the priors are $(1-\eta) \sim$ Beta(4,9); $\beta \sim$ Beta(99,2). Beta distributions are convenient because they limit the range for these two parameters and are easy to draw from. In fact, if $x \sim \chi^2(2a); y \sim \chi^2(2b)$ then $z = x/(x+y) \sim$ Beta(a,b). Since the mean of a Beta(a,b) is $(a/a+b)$ and the variance is $ab/[(a+b)^2(a+b+1)]$, the prior mean of $1-\eta$ is about 0.31 and the prior mean of β about 0.99. The variances, approximately 0.011 and 0.0002, imply sufficiently loose prior distributions.

We draw 10000 replications. Given $1-\eta_0 = 0.55, \beta_0 = 0.97$, we produce candidates for $\theta = [1-\eta, \beta]$ using a reflecting random walk process, i.e. $\theta^\ddagger = \bar\theta + (\theta^{l-1} - \bar\theta) + v_\theta^l$ where $\theta^\ddagger$ is a candidate, θ^{l-1} is the previous draw, $\bar\theta$ is the mean of the process and v_θ^l is a vector of errors. The first component of v_θ (corresponding to $1-\eta$) is drawn from a $\mathbb{U}(-0.03, 0.03)$ and the second (corresponding to β) from a $\mathbb{U}(-0.01, 0.01)$. These ranges produces an acceptance rate of about 75%.

Since we are interested in $(1-\eta)$ and β, we are free to choose which equations to use to estimate them. We arbitrarily choose the ones determining consumption and interest rates. Since these equations are static, Kalman filter estimates of the prediction error are identical to those obtained with OLS equation by equation. We calculate the posterior kernel at both the current $\theta^\ddagger$ and the previous θ^{l-1}, evaluating separately the prior and the likelihood (which is normal in shape). Since $g(1-\eta, \beta) = g(1-\eta)g(\beta)$, we calculate the prior at the draw for each of the two parameters separately. Since the transition matrix is symmetric, the ratio of the kernels at $\theta^\ddagger$ and θ^{l-1} is all that is needed to accept or reject the candidates.

We discard the first 5000 draws and keep the last 5000 for inference. Because of the serial correlation present in the draws, we keep one out of five draws. Therefore, we have a total of 1000 draws for calculating marginal densities and the statistics of interest. We

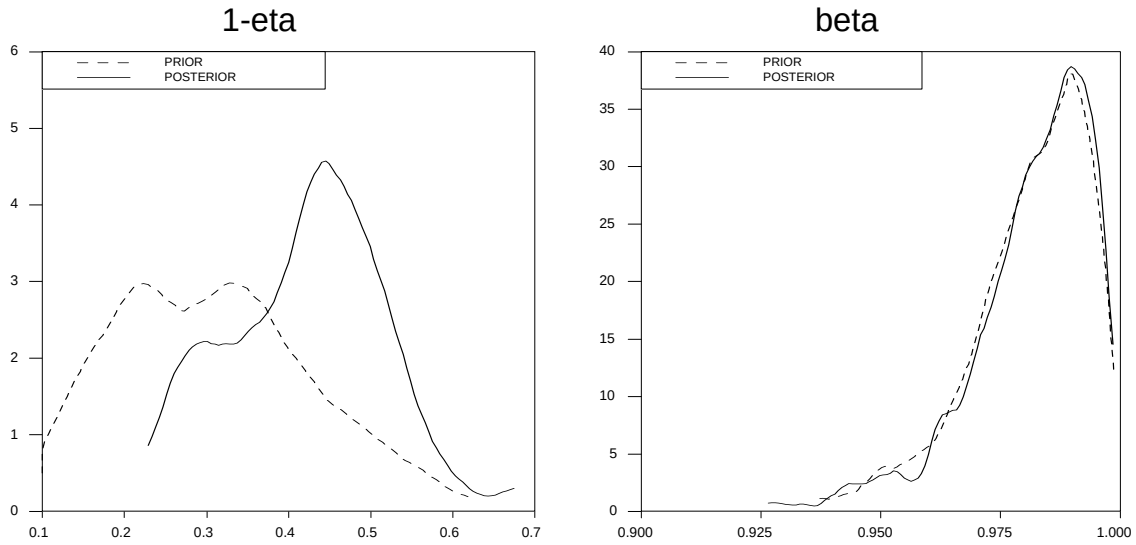


Figure 11.2: Priors and Posteriors, Basic RBC.

check for convergence of the Metropolis algorithm in two ways: splitting the sequences of draws in two and computing a normal test and calculating recursive means for the estimates for each parameter. Convergence was achieved after about 2000 draws.

Figure 11.2 presents the marginal histograms of $1-\eta$ *and* β*, estimated using 1000 skipped draws from the prior and posterior. Two features of the figure are worth mentioning. First, the data is more informative about* $1-\eta$ *than* β*. Second, the posteriors for the two parameters are unimodal and roughly centered around the true parameter values.*

	True	*Posterior 68% range*
var(c)	*40.16*	*[3.65, 5.10e+10]*
var(r)	*1.29e-05*	*[2.55e-04, 136.11]*
cov(c,r)	*-0.0092*	*[-0.15e-05, -0.011]*

Table 11.2: Variances and Covariances

Using the 1000 skipped posterior draws we have calculated three statistics, the variances of consumption and interest rates and the covariance between the two, and compared the posterior 68% credible range with the statistics computed using the "true" parameter values. Table 11.2 shows that the posterior 68% range includes the actual value of the consumption variance but not the one for the real rate or for the covariance. Also, there are posterior combinations of parameters which make the two variances large.

Example 11.10 *In this example we simulated data from an RBC model with habit persistence in consumption and still assume that capital depreciates in one period and leisure does not enter the utility function. We assume* $u(c_t, c_{t-1}) = \ln(c_t - \gamma c_{t-1})$*, set* $\gamma = 0.8$ *and add to the solution the same measurement errors used in equations (11.25)-(11.28). We are interested in knowing how the posterior distributions of* β *and* $1 - \eta$ *look like when we mistakenly assume that there is no habit (i.e. we condition on* $\gamma = 0$*). This experiment is interesting since it can give some indications of the consequences of using a dogmatic (and wrong) prior on some of the parameters of the model.*

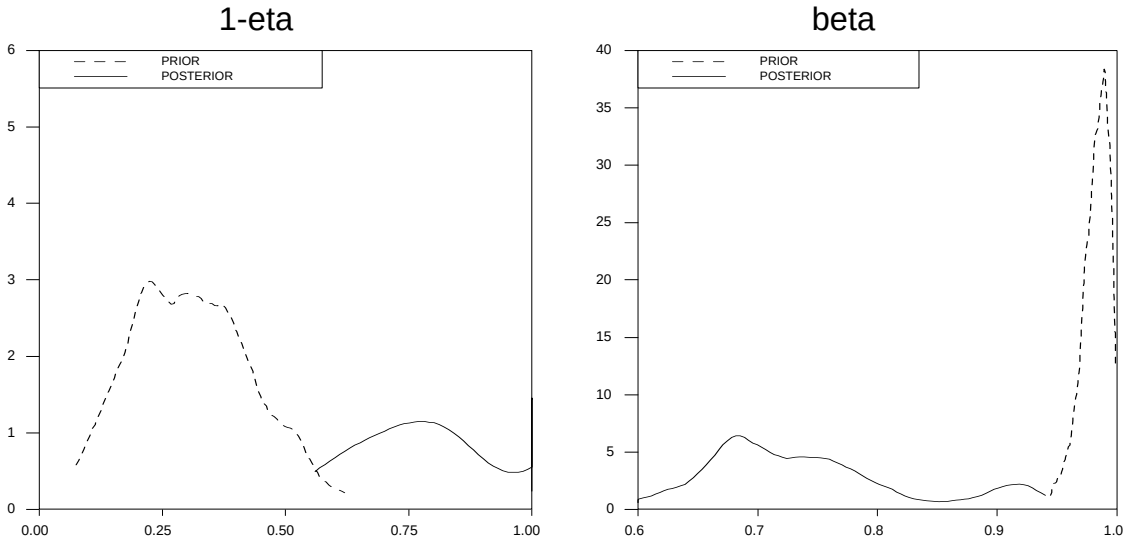


Figure 11.3: Priors and Posteriors, RBC with habit.

Perhaps unsurprisingly, the posterior distributions presented in figure 11.3 are very different from those in figure 11.2. What is somewhat unexpected is that the misspecification is so large that the posterior probability for the "true" parameters is roughly zero.

Exercise 11.27 *Suppose you simulated data from a model where the production function is* $f(K_t, ku_t, \zeta_t) = (K_t ku_t)^{1-\eta}\zeta_t$ *and that the depreciation rate is a function of the utilization, i.e.* $\delta(ku_t) = \delta_0 + \delta_1 ku_t^{\delta_2}$*. Suppose you mistakenly neglect utilization and estimate a model like the one in equations (11.25)-(11.28). Evaluate the distortions induced by this misspecification.*

Example 11.11 *The next example considers a standard New-Keynesian model with sticky prices and monopolistic competition. Our task here is two-fold. First, we want to know how good is this model relative to, say, an unrestricted VAR, in capturing the dynamics of interest rates, the output gap and inflation. Second, we are interested in knowing the*

location of the posterior distribution of some important structural parameters. For example, we would like to know how much price stickiness is needed to match actual dynamics, whether policy inertia is an important ingredient to characterize the data and whether the model has some internal propagation mechanism or if, instead, it relies entirely on the dynamics of the exogenous variables to match the dynamics of the data.

The model economy we use is a simplified version of the structure considered in chapter 2 and is composed of a log-linearized (around the steady state) Euler equation, of a New-Keynesian Phillips curve and of a Taylor rule. We assume that, in equilibrium, consumption is equal to output and use output in deviation from steady states in the Euler equation directly. Each equation has a shock attached to it: there is an iid policy shock, ϵ_{3t}; a cost post shock in the Phillips curve, ϵ_{2t} and an arbitrary demand shock in the Euler equation, ϵ_{4t}. While the latter shock is not necessary for the estimation, it is clearly needed to match the complexity of the process of output, inflation and interest rates observed in the real world. The equations are:

$$gap_t = E_t gap_{t+1} - \frac{1}{\varphi}(i_t - E_t\pi_{t+1}) + \epsilon_{4t} \tag{11.29}$$

$$\pi_t = \beta E_t \pi_{t+1} + \kappa\ gap_t + \epsilon_{2t} \tag{11.30}$$

$$i_t = \phi_r i_{t-1} + (1-\phi_r)(\phi_\pi \pi_{t-1} + \phi_{gap} gap_{t-1}) + \epsilon_{3t} \tag{11.31}$$

where i_t is the nominal interest rate, π_t is the inflation rate, gap_t is the output gap, $\kappa = \frac{(1-\zeta_p)(1-\beta\zeta_p)(\varphi+\vartheta_N)}{\zeta_p}$, where ζ_p is the degree of stickiness in the Calvo setting, β is the discount factor, φ is risk aversion parameter, ϑ_N measures the inverse elasticity of labor supply, ϕ_r the persistence of the nominal rate, while ϕ_π and ϕ_{gap} measure the responses of interest rates to lagged inflation and lagged output gap movements. We assume that ϵ_{4t} and ϵ_{2t} are AR(1) processes with persistence ρ_4, ρ_2 and variances σ_4^2, σ_2^2 while ϵ_{3t} is iid $(0, \sigma_3^2)$.

The model has 12 parameters, $\theta = (\beta, \varphi, \vartheta, \zeta_p, \phi_\pi, \phi_{gap}, \phi_r, \rho_2, \rho_4, \sigma_2^2, \sigma_3^2, \sigma_4^2)$, seven structural and five auxiliary ones, whose posterior distributions need to be found. Our interest centers in the posterior distributions of $(\zeta_p, \phi_r, \rho_2, \rho_4)$. We use US quarterly detrended data from 1948:1 to 2002:1 to estimate the model. We assume that $g(\theta) = \prod_{j=1}^{12} g(\theta_j)$ and use the following priors: $\beta \sim Beta(98,3), \varphi \sim \mathbb{N}(1, 0.375^2), \vartheta_N \sim \mathbb{N}(2, 0.75^2),\ \zeta_p \sim Beta(9,3), \phi_r \sim Beta(6,2), \phi_\pi \sim Beta(1.7, 0.1^2),\ \ \phi_{gap} \sim \mathbb{N}(0.5, 0.05^2), \rho_4 \sim Beta(17,3), \rho_2 \sim Beta(17,3)$ $\sigma_i^{-2} \sim \mathbb{G}(4, 0.1), i = 2,3,4$.

To generate candidate vectors $\theta^\ddagger$, we use a random walk MH algorithm with small uniform errors (tuned up for each parameter so as to achieve a 40% acceptance rate) and check convergence using CUMSUM statistics: $\frac{1}{j}\sum_j \frac{\theta_j^i - E(\theta_j^i))}{\sqrt{var\theta_j^i}}$, where $j = 1, 2, \ldots, J * L + \bar{L}$. Figure 11.4, which presents this statistic, indicates that the chain has converged, roughly, after 15000 draws. Convergence is hard to achieve for ϕ_π and ϕ_{gap}, while it is quickly achieved (at times in less than 10000 iterations) for the other parameters. Note that difficulties with ϕ_π and ϕ_{gap} are not necessarily due to subsample instability (as we will show later). Instead they appear to be related to the non-identifiability of these parameters. Figure 11.5 presents prior and posterior distributions (estimated with kernel methods) using the

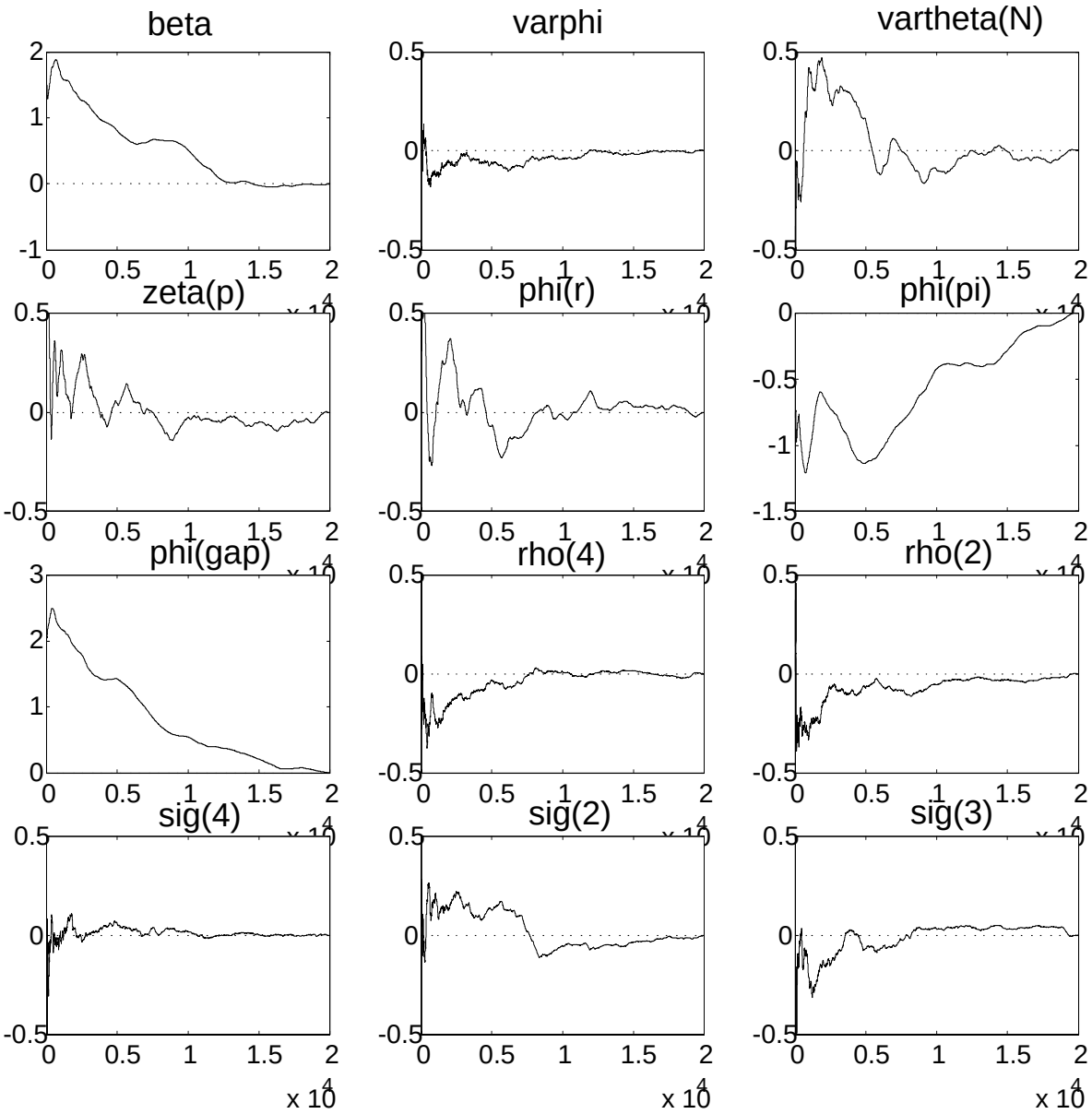


Figure 11.4: CUMSUM statistic.

last 5000 draws. The data appears to be informative in at least two senses. First, in many cases posterior distributions have smaller dispersions than prior ones. Second, in a couple of cases, the whole posterior distribution is shifted relative to the prior. Table 11.3, which presents some statistics of the prior and the posterior, confirms these visual impressions. Note also that, except in a few cases, posterior distributions are roughly symmetric (mean and median coincide).

	Prior		Posterior 1948-2002					Posterior 1948-1981		Posterior 1982-2002	
	mean	std	median	mean	std	min	max	mean	std	mean	std
β	0.98	0.01	0.978	0.976	0.007	0.952	0.991	0.986	0.008	0.983	0.008
φ	0.99	0.37	0.836	0.841	0.118	0.475	1.214	1.484	0.378	1.454	0.551
ϑ_l	2.02	0.75	1.813	2.024	0.865	0.385	4.838	2.587	0.849	2.372	0.704
ζ_p	0.75	0.12	0.502	0.536	0.247	0.030	0.993	0.566	0.200	0.657	0.234
ϕ_r	0.77	0.14	0.704	0.666	0.181	0.123	0.992	0.582	0.169	0.695	0.171
ϕ_π	1.69	0.10	1.920	1.945	0.167	1.568	2.361	2.134	0.221	1.925	0.336
ϕ_{gap}	0.49	0.05	0.297	0.305	0.047	0.215	0.410	0.972	0.119	0.758	0.068
ρ_4	0.86	0.07	0.858	0.857	0.038	0.760	0.942	0.835	0.036	0.833	0.036
ρ_2	0.86	0.07	0.842	0.844	0.036	0.753	0.952	0.831	0.036	0.832	0.036
σ_4	0.017	0.01	0.017	0.017	0.007	0.001	0.035	0.017	0.006	0.016	0.007
σ_2	0.016	0.01	0.011	0.012	0.008	0.0002	0.036	0.016	0.006	0.016	0.007
σ_3	0.017	0.01	0.015	0.016	0.007	0.001	0.035	0.013	0.007	0.014	0.007

Table 11.3: Prior and posterior statistics.

As far as the posterior of the four parameters of interest, note that the shocks are persistent (posterior mean is 0.85) but no pile-up of the posterior distribution for the AR parameters around one occurs. This means that although the model does not have sufficient internal propagation to replicate the dynamics of the data, no exogenous unit-root like processes are needed. (This would have changed if, in the policy rule, interest rates react to output gap and inflation contemporaneously).

The posterior distribution of economic parameters is reasonably centered. The posterior mean of ζ_p, the parameter regulating the stickiness in prices, is only 0.5 implying about two quarters average time between price changes while ϕ_r, the parameter measuring policy persistence, has a posterior mean of 0.7, implying some degree of policy smoothness. However, since the posterior of ζ_p is bimodal, care must be exercised in using the posterior mean as a measure of location. It is important to stress that the two modes are not produced because the prior and the likelihood have different central tendencies.

Finally, the posterior mean of κ is about 3.0, implying a strong reaction of inflation to movements in the output gap, contrary, for example, to the estimates we obtained in chapter 5. Given that the posterior for β has a mean of 0.98, a shock that moves the output gap by one percent implies a long run change in inflation of about 30 percent.

The majority of these conclusions remain splitting the sample in two. For example, ζ_p has a posterior mean of 0.5667 in the 1948-1981 sample and a posterior mean of 0.6573 in the 1982-2002 sample. However, since the posterior standard error is around 0.22, differ-

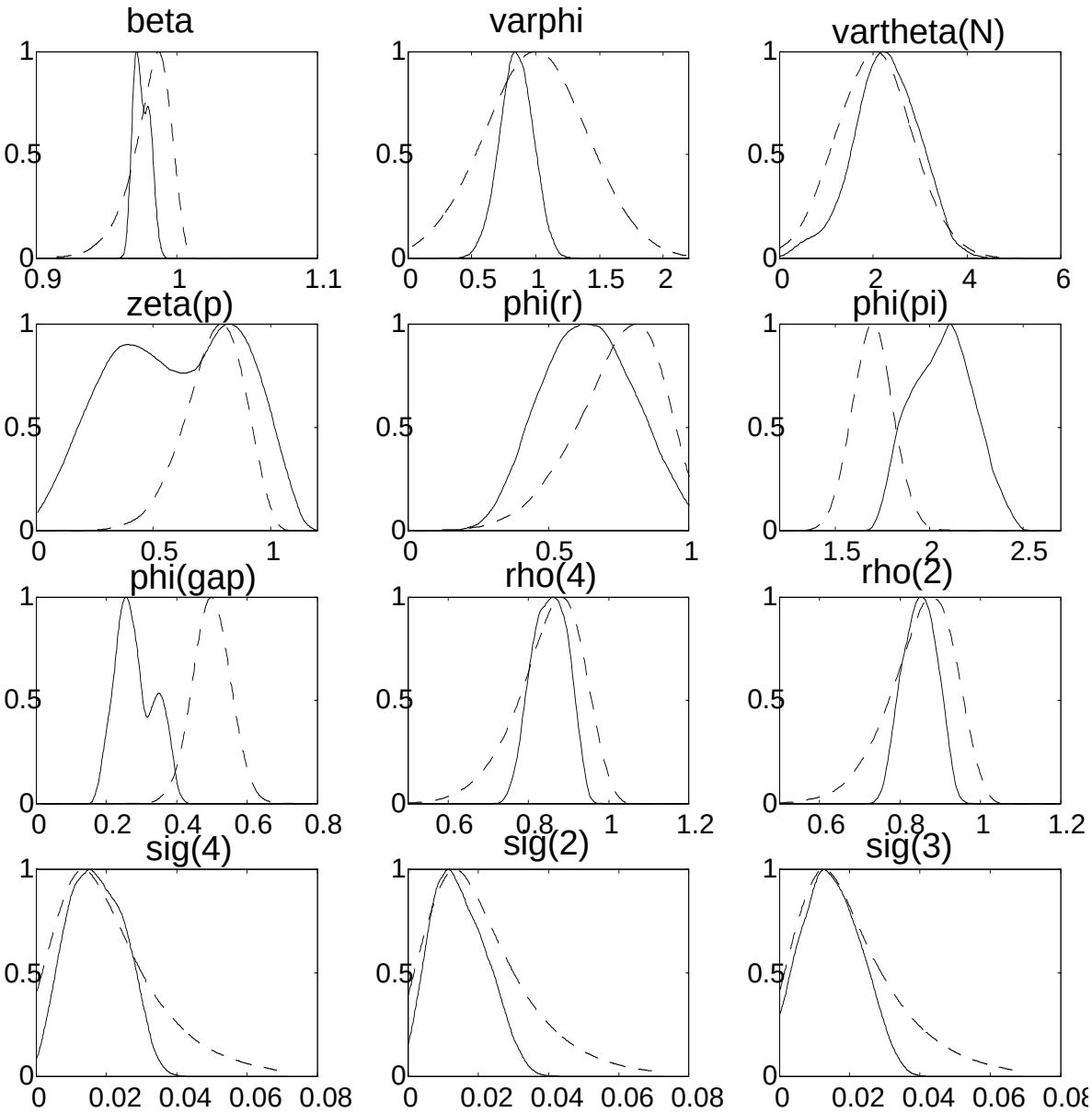


Figure 11.5: Priors (dotted) and Posteriors (solid), Sticky price model.

ences in the two samples are statistically small. Also the other parameters seems to have stable posterior across samples. In particular, splitting the sample does not change the fact that the coefficients in the policy rule imply a strong reaction of interest rates to inflation.

The location and the shape of the posterior distributions are largely independent of the priors we have selected since priors are broadly non-informative. For example, reweighing the posterior draws with a prior whose range is 90% of the range of the original priors in all 12 dimensions produced posterior distributions which are qualitatively very similar to those in figure 11.5.

Finally, we examine the forecasting performance of the model by comparing its predictive density to the one of a VAR(3) and of a BVAR(3) with Minnesota prior and standard parameters (tightness=0.1, linear lag decay and weight on other variables equal 0.5), both with a constant. Bayes factors are small (of the order of 0.19) in both cases indicating that the model can be improved upon in a forecasting sense. Note that, while both alternatives are more densely parametrized than our DSGE model (30 vs. 12 parameters), Bayes factors take model size into account and no adjustment for the number of parameters is needed.

Exercise 11.28 *Consider the model of example 11.11, but replace the Phillips curve with* $\pi_t = \frac{\omega}{1+\omega\beta}\pi_{t-1} + \frac{\beta}{1+\omega\beta}E_t\pi_{t+1} + \frac{\kappa}{1+\omega\beta}\ gap_t + \epsilon_{2t}$ *where* ω *is the degree of indexation of prices. Estimate this model and test if indexation is necessary to match US data.*

Exercise 11.29 *Consider adding to the model of example 11.11 sticky wages using the following optimal wage equation* $\Delta w_t = \beta E_t \Delta w_{t+1} + \frac{(1-\zeta_w)(1-\zeta_w\beta)}{\zeta_w(1+\varsigma_w\vartheta_N)}(mrs_t - (w_t - p_t)) + \epsilon_{2t}$ *where* ζ_w *is the probability of not changing the wage,* ς_w *is the elasticity of substitution between types of labor in production and* mrs_t *is the marginal rate of substitution. Estimate this model and test whether wage stickiness add to the fit of the basic sticky price model.*

11.4.1 A few applied tips

Although the models we have considered so far are small, it has become standard in Central Banks and international institutions to estimate large scale DSGE models with Bayesian methods. Care however should be exercised in such an enterprise for three reasons.

First, many parameters may not be identifiable from the data. For example, if a log-linearized model in deviations from the steady state is used, parameters which enter only in the steady state can never be identified from the data. Less trivially, there are situations when the likelihood function is informative about certain parameters when all the equations of the model are used but uninformative if only a subset of the equations are used to estimate the parameters. At times, it may also happen that only a subset of equations is informative about one parameter and this subset is not necessarily the one an investigator would like to use for inference (an example of this was given in chapter 6). In general, extreme care should be used because informative posterior distributions can be obtained even when parameters are not identified in the data, if a tight enough a-priori distribution is specified. Since in this case the prior and the posterior lie on top of each other, altering both the shape and the spread of the prior may help to detect identification problems.

Second, as we have seen in chapter 6, the likelihood function of a DSGE model may have very large flat sections or very rocky appearance in some dimensions. Once again, if the prior distribution for the parameters is tight, the posteriors appear to be well behaved only because the prior has selected a particular region of the parameter space. While the prior should be used to exclude regions of the parameter space which are unreasonable from an economic point of view, it should also be made reasonably uninformative on the interesting portions of the parameters space to avoid misleading conclusions. Note that multiple peaks in the likelihood may indicate the presence of breaks or multiple regimes and may give important information about the phenomena one is interested in examining. Once again, robustness analysis may inform the investigator on the coherence of the model to the data.

Finally, while it is common to start from a model with a large number of frictions and shocks, Bayesian methods can be used even if the model is grossly misspecified in its dynamics or its probabilistic nature. This means that the type of sequential exercises performed in early calibrated models (e.g. start from a competitive structure with only technology shocks, add government shocks or introduce non-competitive markets, etc.) can be fruitfully employed also here. Frictions and shocks which add little to the ability of the model to reproduce interesting features of the data should be discarded. This analysis could also help to give some of the black-box shocks estimated in the factor literature an interesting economic content.

11.4.2 Comparing models and the data

While Bayesian estimation of structural parameters is simple, it is less straightforward to compare the model outcomes to the data and to assess the superiority of a model among alternative candidate specifications. Two methods are available. The first, preferred by macroeconomists, is based on informal analysis of some interesting economic statistics.

Example 11.12 *Continuing with example 11.11 we present 68 percent impulse response bands to interest rate shocks in figure 11.6.*

While responses are economically reasonable there are three features of the figure which stand out. First, the persistence of the responses is minimal - responses die out after few periods. Second, the responses of inflation and the output gap to an interest rate shocks are negative. Third, despite the assumed price stickiness, the largest inflation effect of an interest rate increase is instantaneous. Figure 11.7 reports response bands obtained estimating the model over different subsamples. The figure is constructed estimating posterior distributions of the parameters, keeping a constant number of observations, but moving the sample over time. It is remarkable that the sign, the shape and the magnitude of the posterior 68% credible band is unchanged as we move from the late 1970's to the early 2000. Hence, the transmission properties of monetary shocks have hardly changed over the last 30 years.

As an alternative to the presentation of economic statistics of various nested or non-nested models , one could compute measures of forecasting performance of various specifications. As we have seen in chapter 9, the predictive density is the product of one-step ahead

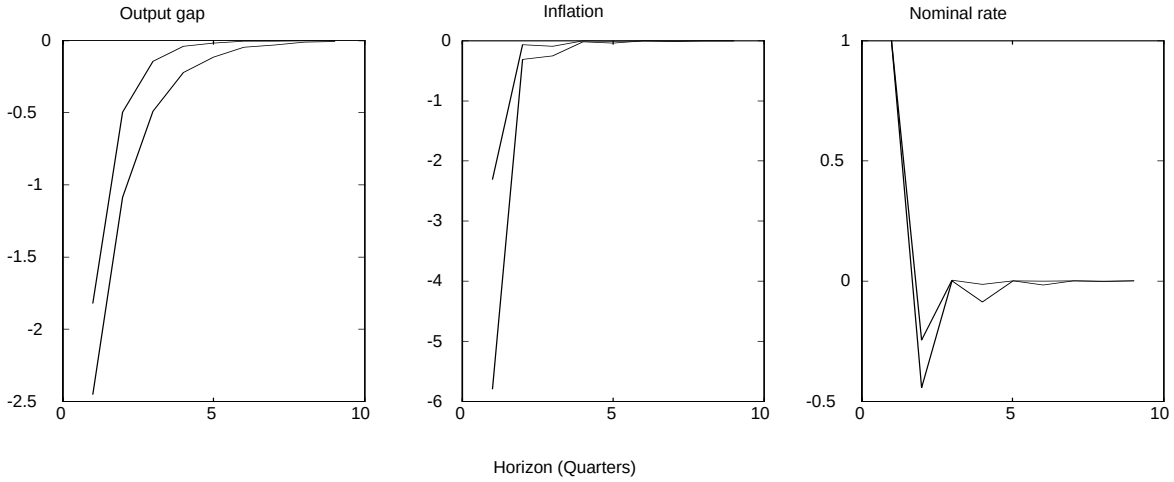


Figure 11.6: Impulse Responses, sample 1948-2002.

forecast errors. Hence, selecting a model using Bayes factors, as we have done in example 11.11, is equivalent to choosing the specification with smallest one-step (in-sample) MSE. Clearly, out-of-sample forecasting races are also possible, in which case predictive Bayes factors can be computed (see e.g De Jong, Ingram and Whiteman (2000)). This is easy to do: we leave it to the reader to work out the details.

Exercise 11.30 *Show how to construct the predictive density of future endogenous variables* $y_{t+\tau},\ \tau = 1, 2, \ldots$ *(Hint: use the restricted VAR representation of a DGSE model).*

Despite their popularity, Bayes factors (or Posterior odds ratios) may not be very informative about the quality of the approximation to the data, in particular, when the models one wishes to compare are misspecified.

Example 11.13 *Suppose that there are three models, two structural ones,* $(\mathcal{M}_1, \mathcal{M}_2)$*, and a reference one* $(\mathcal{M}_3)$*, densely parametrized (e.g. a VAR). The Bayes factor between the two structural models is* $\frac{f(y,\mathcal{M}_1)}{f(y)} \times \frac{f(y)}{f(y,\mathcal{M}_2)}$ *where* $f(y) = \int f(y, \mathcal{M}_i) d\mathcal{M}_i$*. If we use a 0-1 loss function, and assume that the prior probability of each model is 0.5, the posterior risk is minimized by selecting* $\mathcal{M}_1$ *if Bayes factor exceeds one. The presence of a third model does not affect the choice since it only enters in the calculation of* $f(y)$*, which cancels out of the Bayes factor. If the prior odds do not depend on this third model, the PO ratio will also be independent of it. When* $\mathcal{M}_1$ *and* $\mathcal{M}_2$ *are misspecified, they will have low posterior probability relative to* $\mathcal{M}_3$ *but this has no influence on inference one makes. Hence, comparing misspecified models with a Bayes factor may be uninteresting: one model may be preferable to another one but it may still have close to zero posterior probability.*

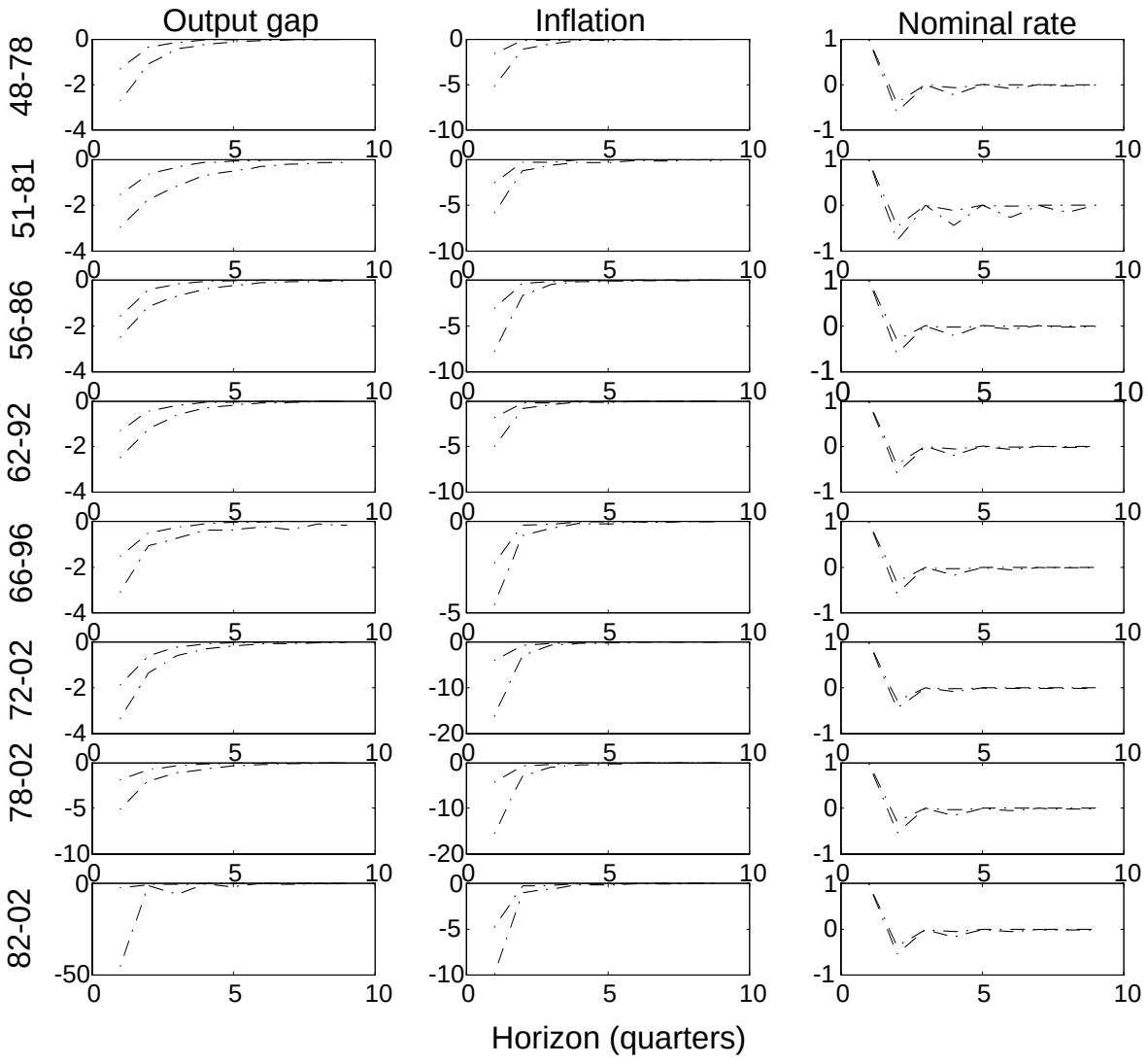


Figure 11.7: Impulse responses, various samples

Schorfheide (2000) provided a simple procedure to choose among misspecified models (in his case a cash-in-advance and a working capital model), both of which are likely to have very low posterior probability. The actual data is assumed to be a mixture of the competing structural models and of a reference one, which has two characteristics: (i) it is more densely parametrized than the DSGE models; (ii) it can be used to compute a vector of population functions $h(\theta)$. One such model could be a VAR or a BVAR. Given this setup, model comparisons can be undertaken using loss functions. In particular, when several models are available, the following evaluation algorithm could be used:

Algorithm 11.5

1) *Compute the posterior distribution for the parameters of each model, using tractable priors and one of the available posterior simulators.*

2) *Obtain the predictive density of the data, for each $\mathcal{M}_i$, that is, compute $f(y|\mathcal{M}_i) = \int f(y|\theta_i, \mathcal{M}_i) g(\theta_i|\mathcal{M}_i) d\theta_i$.*

3) *Compute posterior probabilities $\tilde{P}_i = \frac{\bar{P}_i f(y|\mathcal{M}_i)}{\sum_i \bar{P}_i f(y|\mathcal{M}_i)}$, where $\bar{P}_i$ is the prior probability of model i. Note that if the distribution of y is degenerated under $\mathcal{M}_i$ (e.g if number of shocks is smaller than the number of endogenous variables), $\tilde{P}_i = 0$.*

4) *Calculate the posterior distribution of $h(\theta)$ for each model and average using posterior probabilities i.e. obtain $g(h(\theta)|y, \mathcal{M}_i)$, and $g(h(\theta)|y) = \sum_i \tilde{P}_i g(h(\theta)|y, \mathcal{M}_i)$. If all but model i' produce degenerate distributions for θ, $g(h(\theta)|y) = g(h(\theta)|y, \mathcal{M}_{i'})$.*

5) *Setup a loss function $\mathsf{L}(h_T, h_i(\theta))$ measuring the discrepancy between model's i predictions and data h_T. Since the optimal predictor in model $\mathcal{M}_i$ is $\hat{h}_i(\theta) = arg\min_{h_i(\theta)} \int \mathsf{L}(h_T, h_i(\theta)) g(h_i(\theta)|y, \mathcal{M}_i) dh_T$, one can compare models using the risk of $\hat{h}_i(\theta)$ under the overall posterior distribution $g(h(\theta)|y_T)$, i.e. $\mathsf{R}(\hat{h}_i(\theta)|y) = \int \mathsf{L}(h_T, \hat{h}_i(\theta)) g(h(\theta)|y) dh_T$.*

In step 5) $\mathsf{R}(\hat{h}_i(\theta)|y_T)$ measures how well model $\mathcal{M}_i$ predicts h_T. Hence a model is preferable to another one if it has a lower risk. Note also that while model comparison is relative, $g(h(\theta)|y_T)$ takes into account information from all models. Taking step 5) further, one should notice that, for each i, θ can be selected so as to minimize $\mathsf{R}(\hat{h}_i(\theta)|y_T)$. Such an estimate provides a lower bound to the posterior risk obtained by the "best" candidate model in the dimensions represented by h_T.

To make algorithm 11.5 operative a loss function must be selected. We have presented a few options in chapter 9. For DSGE models, the most useful are:

(a) Quadratic loss: $\mathsf{L}_2(h_T, h(\theta)) = (h_T - h(\theta))' W (h_T - h(\theta))$; W is an arbitrary positive definite weighting matrix.
(b) Penalized loss: $\mathsf{L}_p(h_T, h(\theta)) = \mathcal{I}_{[g(h(\theta)|y) < g(h_T|y)]}$ where $\mathcal{I}_{[x_1 < x_2]} = 1$ if $x_1 < x_2$.
(c) χ^2 loss: $\mathsf{L}_{\chi^2}(h(\theta), h_T) = \mathcal{I}_{[\mathcal{Q}_{\chi^2}(h(\theta)|y) > \mathcal{Q}_{\chi^2}(h_T|y)]}$, $\mathcal{Q}_{\chi^2}(h(\theta)|y) = (h(\theta) - E(h(\theta)|y))'$

$\Sigma_{h(\theta)}^{-1}(h(\theta) - E(h(\theta)|y))$, where $\Sigma_{h(\theta)}$ is the covariance of $h(\theta)$ and $\mathcal{I}_{[x_1>x_2]} = 1$ if $x_1 > x_2$.
(d) 0/1 loss: $\mathsf{L}(h_T, h(\theta), \epsilon) = 1 - \mathcal{I}_{\epsilon(h(\theta)}(h_T)$, where $\epsilon(h(\theta))$ is a ϵ-neighborhood of $h(\theta)$.

Three features of these loss functions should be mentioned. First, with penalized and χ^2 loss functions two DSGE models are compared on the basis of the height of the posterior distribution at $h_i(\theta)$. Second, with a quadratic loss function comparison is based on the weighted distance between $h_i(\theta)$ and the posterior mean. Third, as mentioned, a 0-1 loss implies that $\mathcal{M}_1$ is preferred if the posterior odds exceeds one.

Exercise 11.31 *(i) Show that* $\mathsf{R}_2 = (h_T - E(h(\theta))|y))'W(h_T - E(h(\theta))|y)) + \varrho_0$ *where* ϱ_0 *does not depend on* $Eh(\theta)$*. How would you choose* W *optimally?*
(ii) Show that if $g(\theta|y)$ *is normal,* $\mathsf{L}_2 = \mathsf{L}_{\chi^2}$ *and that the optimal predictor is* $E(h(\theta)|y, \mathcal{M}_i)$*.*
(iii) Verify that the optimal predictor for the L_p *loss is the mode of* $g(h(\theta)|y, \mathcal{M}_i)$*.*

Two interesting special cases obtain when the L_2 loss is used.

Exercise 11.32 *(Schorfheide) Suppose that there are three models.*
(i) Suppose $\tilde{P}_1 \xrightarrow{P} 1$*,* $E(h_i(\theta)|y_T, \mathcal{M}_i) \xrightarrow{P} \bar{h}_i(\theta)$ *and* $\bar{h}_1(\theta) - \bar{h}_2(\theta) = \delta_\theta$*, where* $|\delta_\theta| > 0$*. Show that as* $T \to \infty$*,* $\mathsf{R}_2(\hat{h}_1(\theta)) \xrightarrow{P} 0$ *and* $\mathsf{R}_2(\hat{h}_2(\theta)) \xrightarrow{P} \delta_\theta' W \delta_\theta$*.*
(ii) Suppose that as $T \to \infty, \tilde{P}_{3,T} \to 1$ *and* $E(h_i(\theta)|y, \mathcal{M}_i) \xrightarrow{P} \bar{h}_i(\theta)$*. Show that* $E(h(\theta)|y) - E(h(\theta)|y, \mathcal{M}_3) \xrightarrow{P} 0$

Exercise 11.32 reaches a couple of interesting conclusions. First, if for any positive definite W, model $\mathcal{M}_1$ is better than $\mathcal{M}_2$ with probability one, model selection using the L_2 loss is consistent and gives the same result as a PO ratio in large samples. To restate this concept in another way, under these conditions L_2-model comparison is based on the relative one-step ahead predictive ability. Second, if the two models are so misspecified that their posterior probability goes to zero as $T \to \infty$, the ranking of these models only depends on the discrepancy between $E(h(\theta)|y, \mathcal{M}_3) \approx E(h(\theta)|y)$ and $\hat{h}_i(\theta), i = 1, 2$. If $\mathcal{M}_3$ is any empirical model, then using a L_2 loss is equivalent to compare sample and population moments obtained from different models. That is to say, an informal comparison between the predictions of the model and the data, as it is done in the simplest calibration exercises, is optimal from a Bayesian point of view when one makes decisions based on the L_2 loss function and the models are highly misspecified. Intuitively, this outcome obtains because the posterior variance of $h(\theta)$ does not affect the ranking of models. Note that this conclusion does not hold with the L_p or the L_{χ^2} loss.

Example 11.14 *Continuing with example 11.11, we calculate the risk associated with the model when* $h(\theta)$ *represents the persistence of inflation and persistence is measured by the height of the spectrum at the zero frequency. This number is large (227.09), reflecting the inability of the model to generate persistence in inflation. In comparison, for example, the risk generated by a univariate AR(1) is 38.09.*

11.4.3 DSGEs and VARs, once again

As mentioned in chapter 10, it is possible to use data simulated from a DSGE model to construct a prior for reduced form VAR coefficients. Such an approach is advantageous since it jointly allows posterior estimation of both reduced form and structural parameters. We have already derived the posterior for VAR parameters in section 2.5 of chapter 10. Here we describe how to obtain posterior distributions for the structural ones. Let $f(y|\alpha, \Sigma_e)$ be the likelihood function of the data, conditional on the VAR parameters, let $g(\alpha, \Sigma_e|\theta)$ be the prior for the VAR parameters, conditional on the DSGE model parameters, and $g(\theta)$ the prior distribution for DSGE parameters. Here $g(\alpha, \Sigma_e|\theta)$ is the prior for the reduced form parameters induced by the prior on the structural parameters and the details of the model. The joint posterior of VAR and structural parameters is $g(\alpha, \Sigma_e, \theta|y) = g(\alpha, \Sigma_e, |\theta, y)g(\theta|y)$.

We have seen that $g(\alpha, \Sigma_e, |\theta, y)$ has a Normal-inverted Wishart form so that it can be easily computed either analytically or by simulation. The computation of $g(\theta|y)$ is slightly more complicated since its form is unknown. The kernel of this distribution is $\breve{g}(\theta|y) = f(y|\theta)g(\theta)$ where

$$f(y|\theta) = \int f(y|\alpha, \Sigma_e)g(\alpha, \Sigma_e, \theta)d\alpha d\Sigma = \frac{f(y|\alpha, \Sigma_e)g(\alpha, \Sigma_e|\theta)}{g(\alpha, \Sigma_e|y, \theta)} \tag{11.32}$$

Since the posteriors of (α, Σ_e) depends on θ only through y, $g(\alpha, \Sigma_e|y, \theta) = g(\alpha, \Sigma_e|y)$ and we can use the fact that both the numerator and the denominator of (11.32) have Normal-Wishart form to obtain

$$\begin{aligned} f(y|\theta) &= \frac{|(X^s)'(\theta)X^s(\theta) + X'X|^{-0.5m}|(T_s + T)\tilde{\Sigma}_e(\theta)|^{-0.5(T_s+T-k)}}{|(X^s)'(\theta)X^s(\theta)|^{-0.5m}|T_s\bar{\Sigma}_e^s(\theta)|^{-0.5(T_s-k)}} \\ &\times \frac{(2\pi)^{-0.5mT}\ 2^{0.5m(T_s+T-k)}\prod_{i=1}^{m}\Gamma(0.5*(T_s+T-k+1-i))}{2^{0.5m(T_1-k)}\prod_{i=1}^{m}\Gamma(0.5*(T_s-k+1-i))} \end{aligned} \tag{11.33}$$

where T_s is the number of observations from the DSGE model added to the actual data, Γ is the Gamma function, $X = (I \otimes \mathsf{X})$ includes all the lags of y, the superscript s indicates simulated data and k is the number of coefficients in each VAR equation.

Exercise 11.33 *Suggest an algorithm to draw sequences from $g(\theta|y)$.*

11.4.4 Non linear specifications

So far we have examined DGSE models which are (log-)linearized around some pivotal point. As seen in chapter 2, there are applications for which (log-)linearizations are unappealing, for example when economic experiments involve changes of regime or large perturbation of the relationships. In these cases one may want to work directly with the nonlinear version of the model and some steps of the algorithms we have presented in this chapter need to be

modified to take this into account. Consider the model

$$y_{2t+1} = h_1(y_{2t}, \epsilon_{1t}, \theta) \tag{11.34}$$

$$y_{1t} = h_2(y_{2t}, \epsilon_{2t}, \theta) \tag{11.35}$$

where ϵ_{2t} are measurement errors, ϵ_{1t} are structural shocks, θ is a vector of structural parameters, y_{2t} is the vector of states and y_{1t} is the vector of controls. Let $y_t = (y_{1t}, y_{2t})$, $\epsilon_t = (\epsilon_{1t}, \epsilon_{2t})$, $y^{t-1} = (y_0, \ldots, y_{t-1})$ and $\epsilon^t = (\epsilon_1, \ldots, \epsilon_t)$. The likelihood of the model is $\mathcal{L}(y^T, \theta y_{20}) = \prod_{t=1}^{T} f(y_t|y^{t-1}, \theta) f(y_{20}, \theta)$. Integrating the initial conditions and the shocks out, the likelihood can be written as (see Fernandez-Villaverde and Rubio-Ramirez (2003))

$$\mathcal{L}(y^T, \theta) = \int [\prod_{t=1}^{T} \int f(y_t|\epsilon^t, y^{t-1}, y_{20}, \theta) f(\epsilon^t|y^{t-1}, y_{20}, \theta) d\epsilon^t] f(y_{20}, \theta) dy_{20} \tag{11.36}$$

where y_{20} is the initial state of the model. Clearly (11.36) is intractable. However, if we have L draws for y_{20} from $f(y_{20}, \theta)$ and L draws for $\epsilon^{t|t-1,l}$, $l = 1, \ldots, L$, $t = 1, \ldots, T$, from $f(\epsilon^t|y^{t-1}, y_{20}, \theta)$ we can approximate (11.36) using

$$\mathcal{L}(y^T, \theta) = \frac{1}{L} [\prod_{t=1}^{T} \frac{1}{L} \sum_{l} f(y_t|\epsilon^{t|t-1,l}, y^{t-1}, y_{20}^l, \theta)] \tag{11.37}$$

Drawing from $f(y_{20}, \theta)$ is simple, but drawing from $f(\epsilon^t|y^{t-1}, y_{20}, \theta)$ is, in general, complicated. Fernandez-Villaverde and Rubio-Ramirez suggest to use $f(\epsilon^{t-1}|y^{t-1}, y_{20}, \theta)$ as importance sampling for $f(\epsilon^t|y^{t-1}, y_{20}, \theta)$ as the next algorithm indicates:

Algorithm 11.6

1) *Draw y_{20}^l from $f(y_{20}, \theta)$. Draw $\epsilon^{t|t-1,l}$ L times from $f(\epsilon^t|y^{t-1}, y_{20}^l, \theta) = f(\epsilon^{t-1}|y^{t-1}, y_{20}^l, \theta) f(\epsilon_t|\theta)$.*

2) *Construct $IR_t^l = \frac{f(y_t|\epsilon^{t|t-1,l}, y^{t-1}, y_{20}^l, \theta)}{\sum_{l=1}^{L} f(y_t|\epsilon^{t|t-1,l}, y^{t-1}, y_{20}^l, \theta)}$ and assign it to each draw $\epsilon^{t|t-1,l}$.*

3) *Resample from $\{\epsilon^{t|t-1,l}\}_{l=1}^{L}$ with probabilities equal to IR_t^l. Call this draw $\epsilon^{t,l}$.*

4) *Repeat steps 1)-3) for every $t = 1, 2, \ldots, T$.*

Step 3) is crucial to make the algorithm work. If it is omitted, only one particle will asymptotically remain and the integral in (11.36) diverges as $T \to \infty$. The resampling step prevents this from happening. Note that such a step is similar to the one employed in genetic algorithms: you resample from candidates which have high probability and create new branches at each step.

Clearly this algorithm is computationally demanding: in fact at each iteration, the model needs to be solved to find an expression for $f(y^t|\epsilon^t, y^{t-1}, y_{20}, \theta)$. At this point only the most basic RBC model has been estimated by non-linear likelihood methods and some gains have been reported by Fernandez-Villaverde and Rubio-Ramirez (2004). When Bayesian analysis is performed algorithm 11.6 must be inserted between steps 3) and 4) of algorithm 11.3. This makes a full Bayesian non-linear approach unfeasible on currently available computers.

11.4.5 Which approach to use?

There is surprisingly little work comparing estimation approaches in models which are misspecified, tightly parametrized and featuring less driving forces than endogenous variables. Ruge Murcia (2002) is one recent example. Despite the lack of formal evidence, there are few general ideas which may be useful to the applied investigator.

First, there are economic and statistical advantages in jointly estimating a system of structural equations. From an economic point of view, this is appealing since parameter estimates are obtained employing all the restrictions implied by the model. On the other hand, statistical efficiency is enhanced when all available information is used. Joint estimation may be problematic when a researcher is not necessarily willing to subscribe all the details of a model. After all, tight parameter estimates which are economically unreasonable are hard to justify and interpret.

Misspecification, a theme we have repeatedly touched upon in several chapters of this book, creates problems to full information estimation techniques in at least two ways. When the number of shocks is smaller than the number of endogenous variables, parameter estimates can be obtained only from a restricted number of series - essentially transforming full information methods into limited information ones. Furthermore, since not all variables have the same informational content about the parameters of interest, one is forced to experiment, with little guidance from economic or statistical theory. Second, if the model can not be considered the DGP (because of the assumptions made or because of the purely qualitative nature of the behavioral relationships it describes) both full information estimation and testing are problematic. Maximum likelihood, in fact, attempts to minimize the largest discrepancy between the model's equations and the data. That is to say, it will chose parameter estimates that are best in the dimensions where misspecification is the largest. Therefore, it is likely to produce estimates which are either unreasonable from an economic point of view or on the boundary of the parameter space.

There are few solutions to these problems. Adding measurement errors may eliminate the singularity of the system but it can not remedy dynamic misspecification problems. Adding serially correlated measurement errors, on the other hand, may solve both problems, but such an approach lacks economic foundation. Roughly speaking, it amounts to giving up the idea that the model is a good representation of the data, both in an economic and in a statistical sense. The methods we have described in the last three chapters can elegantly deal with these problems. The prior plays the role of a penalty function and if it is appropriately specified it may make a full information approach look for a local, but economically interesting maximum of the problem. In addition, it may reduce both biases and skewness in ML estimates. However, it is still to be proved that computer intensive MCMC methods have good size and power properties in the types of models we have studied in this book. The simple example we have conducted in this chapter indicates that a lot more work is needed.

The alternative is to use less information and therefore be theoretically less demanding on the quality of the approximation of the model to the data. Still, the singularity of the system imposes restrictions on the vector of moments (functions) used to estimate the struc-

tural parameters - the functions must be linearly independent, otherwise the asymptotic covariance matrix of the estimates will not be well defined. Nevertheless, there are situations when the model is extremely singular (for example, there is one source of shocks and ten endogenous variables) and limited information procedures like GMM, SMM or indirect inference may paradoxically use more information than ML. We have also mentioned that limited information approaches may fall into logical inconsistencies whenever they claim to approximate only parts of the DGP. To avoid these inconsistencies, what an investigator wants to explain and "the rest" should naturally have a block recursive structure, which is hardly a feature of currently available DSGE models.

Despite the remarkable progress in the specification of DSGE models, one may still prefer to take the point of view that models are still too stylized to credibly represent the data and choose an estimation approach where only the qualitative implications (as opposed to the quantitative ones) are entertained. Such an approach sidesteps both the singularity and the misspecification issues, since qualitative implications can be embedded, as seen in chapter 4, as identification devices for structural VAR models. Interacting DGSE and VAR models either informally or more formally, as in Del Negro and Schorfheide (2004), seems to be the most promising way to bring stylized models onto the data.

In terms of computations, a VAR based approach has clear advantages. Bayesian and ML estimation are time consuming especially when the objective function is not well behaved (a typical case with DSGE models), while SMM and Indirect Inference may require substantial computer capabilities. GMM is a close competitor, but its severe small sample problems may well wipe out the gains from simplicity. In particular, the large biases we discussed in chapter 5 may make GMM (and potentially SMM) unsuitable for macroeconomic problems where samples are typically short and when they are not, breaks or regime changes make the time series of data heterogeneous.

It is also important to stress that different small sample distributions for the structural parameters do not necessarily translate in statistical and economically large differences in interesting functions a researcher wants to compute. For example, Ruge-Murcia (2002) documents that ML, GMM, SMM and Indirect Inference have somewhat different small sample biases and markedly different efficiency properties. Yet, small sample impulse response bands computed with estimates obtained with the four approaches are similar in size and shape.

TECHNICAL APPENDIX TO:
Price Dispersion in Monetary Unions the Role Of Fiscal Shocks
by F. Canova and E. Pappa

Intended for Referees Information and not for Publication

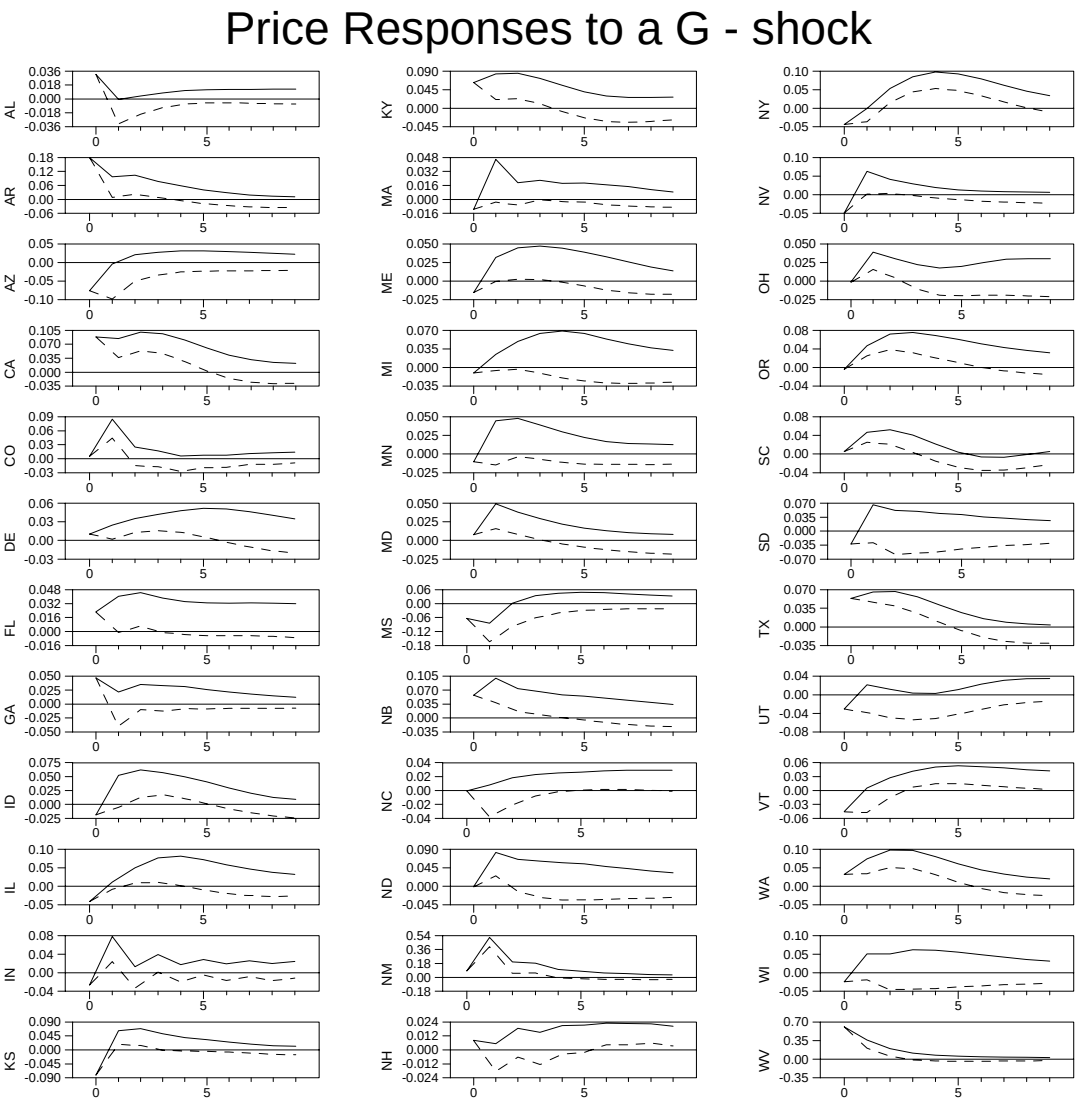


Figure A1: US States

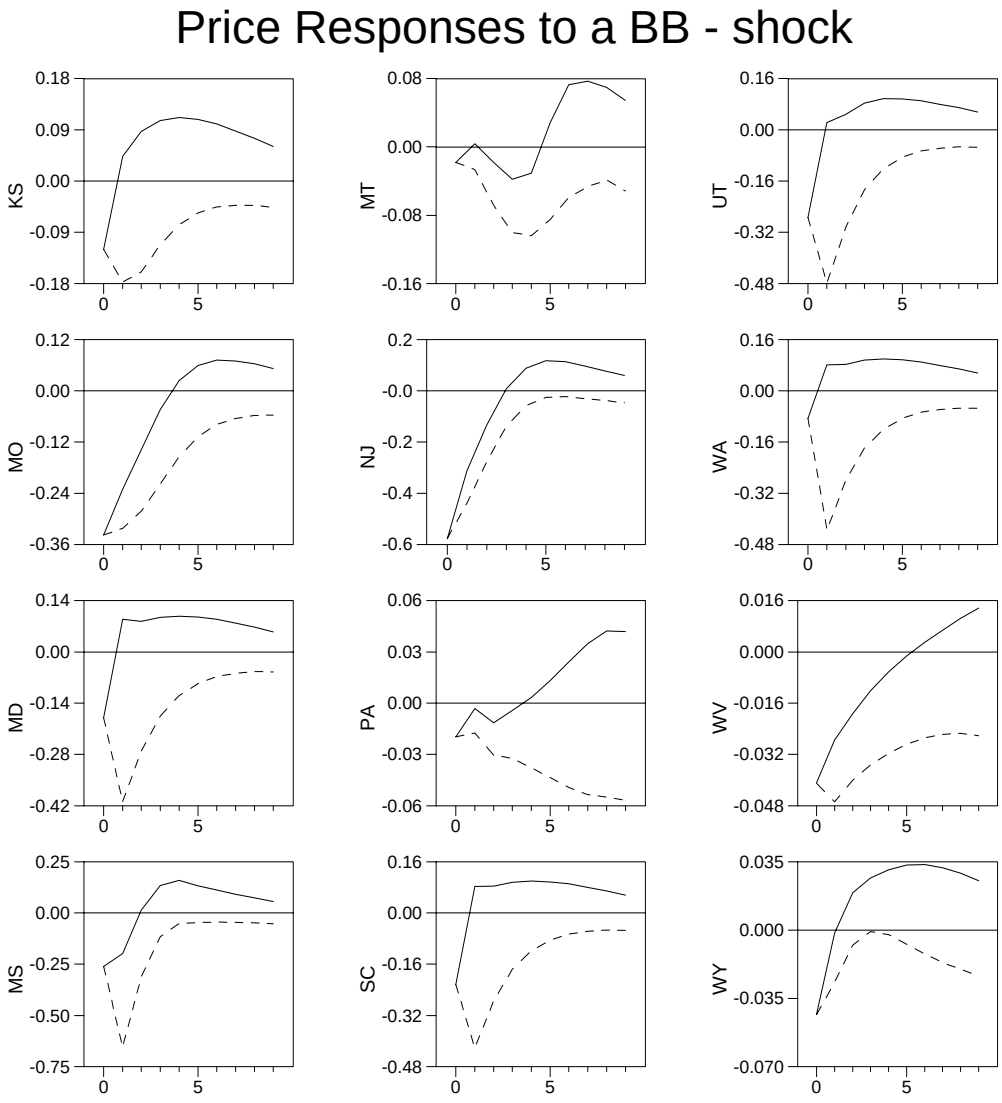


Figure A2: US States

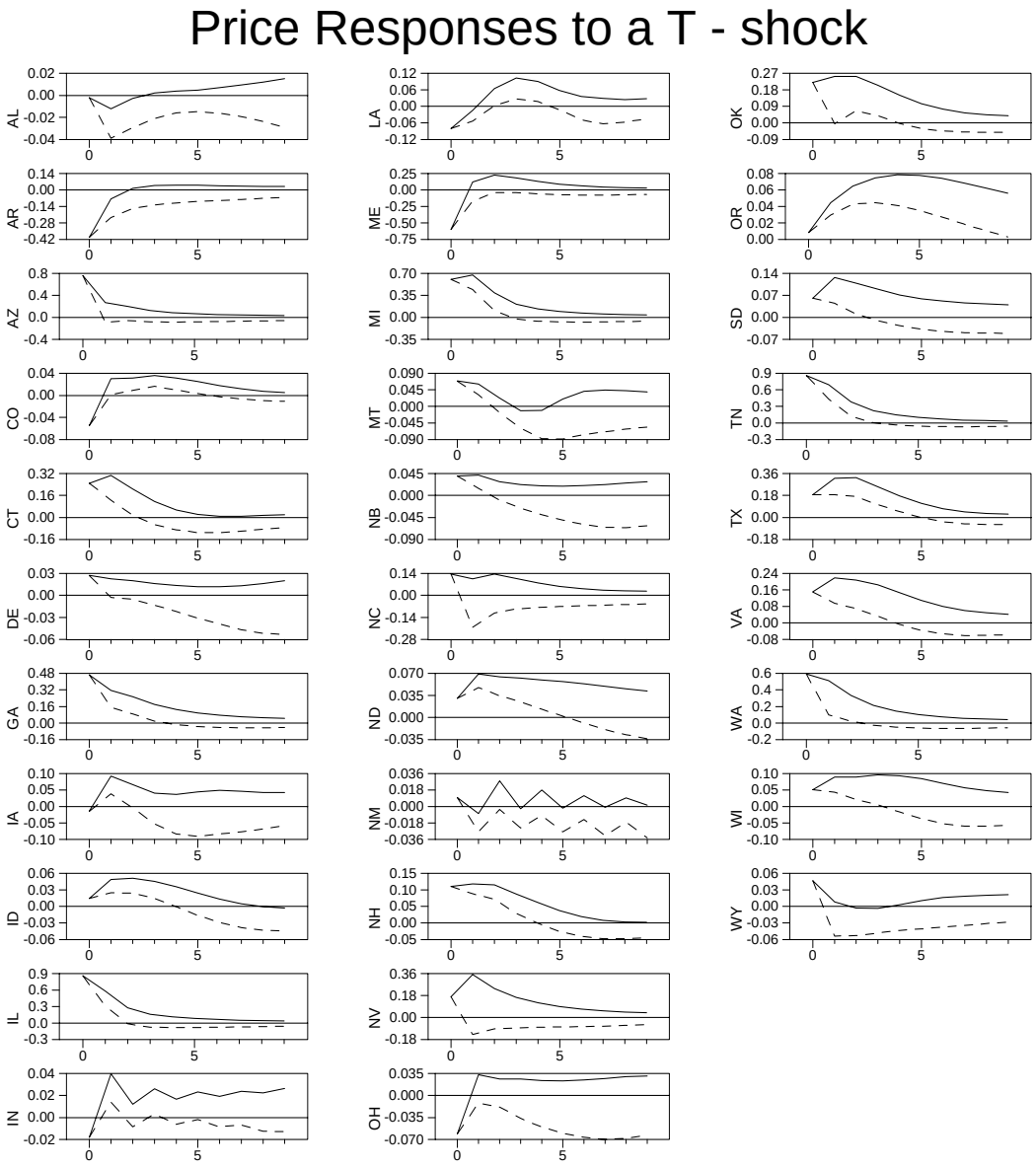


Figure A3: US States

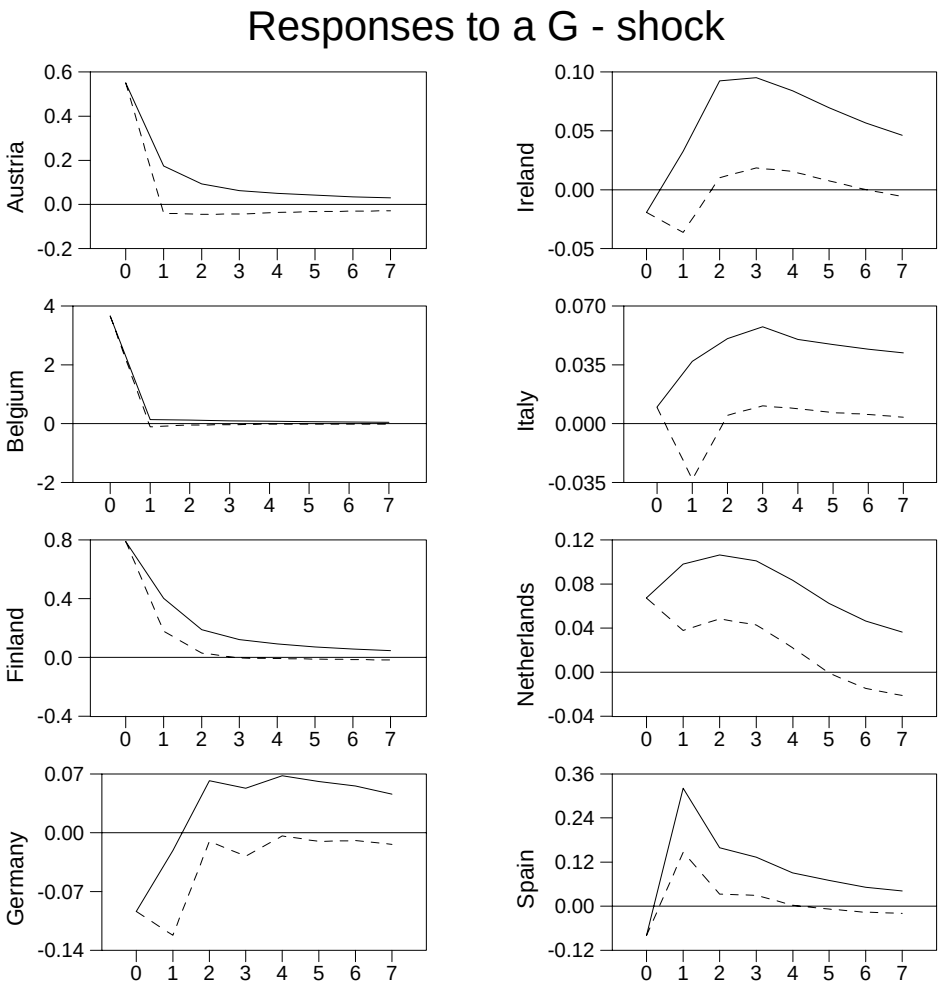


Figure A4: EMU

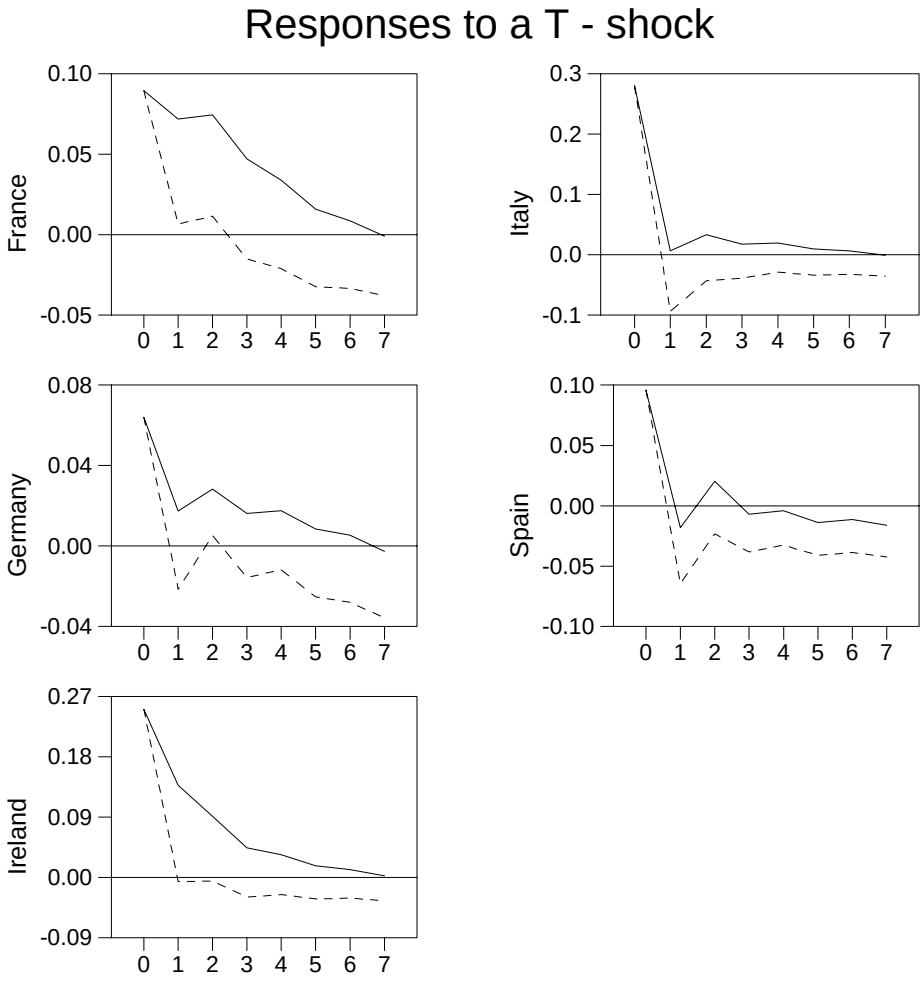


Figure A5: EMU

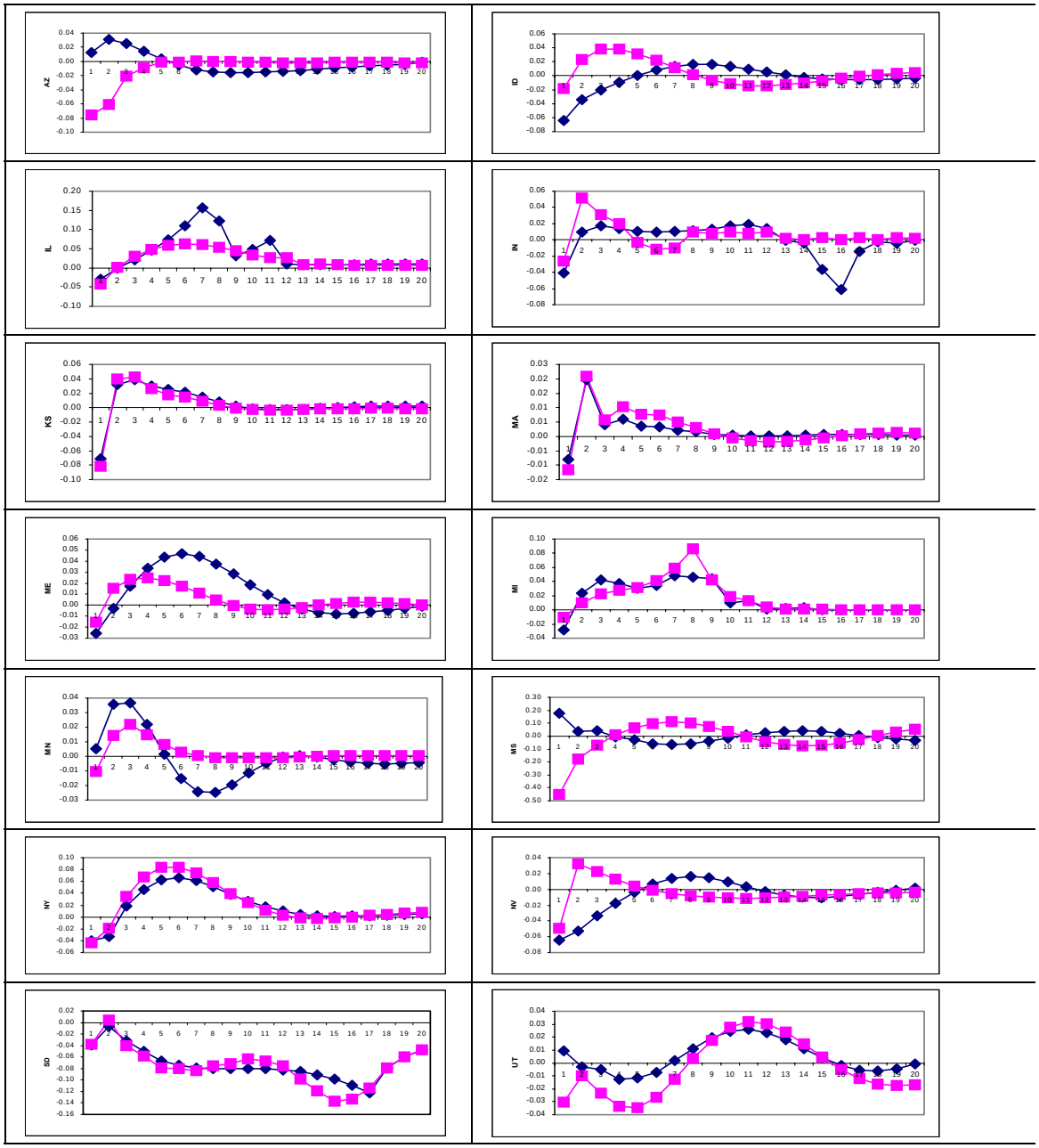


Figure A.6 Spillovers of G shocks (Blue Regional Prices, Red Local Prices)

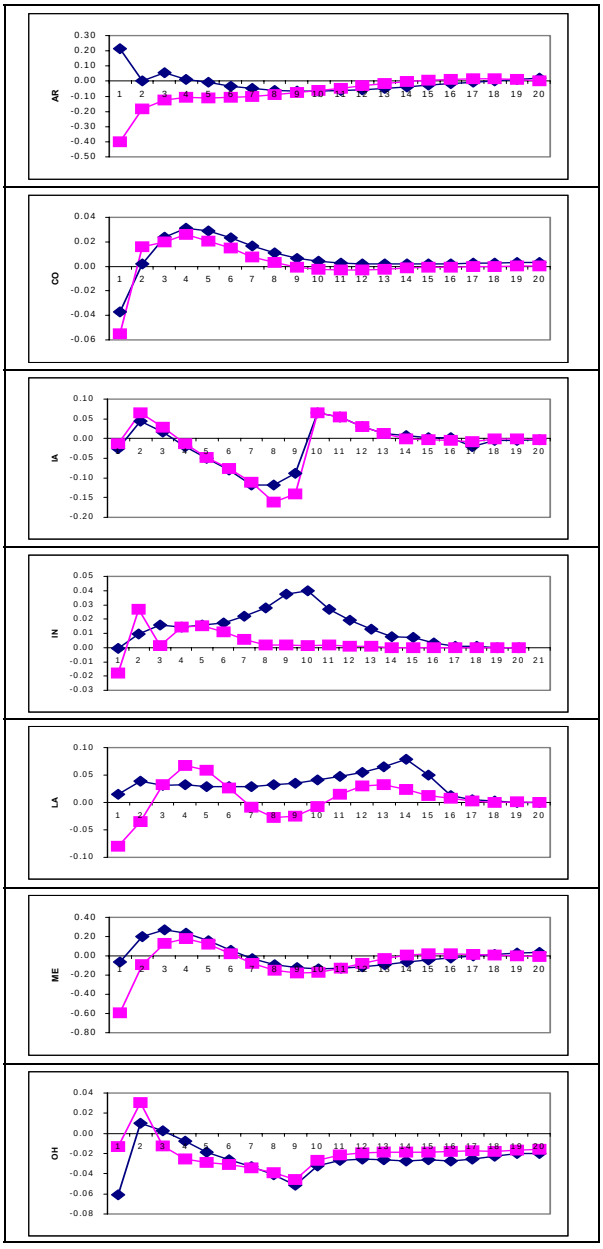


Figure A.7: Spillovers of T Shocks (Blue Regional Prices, Red Local Prices)

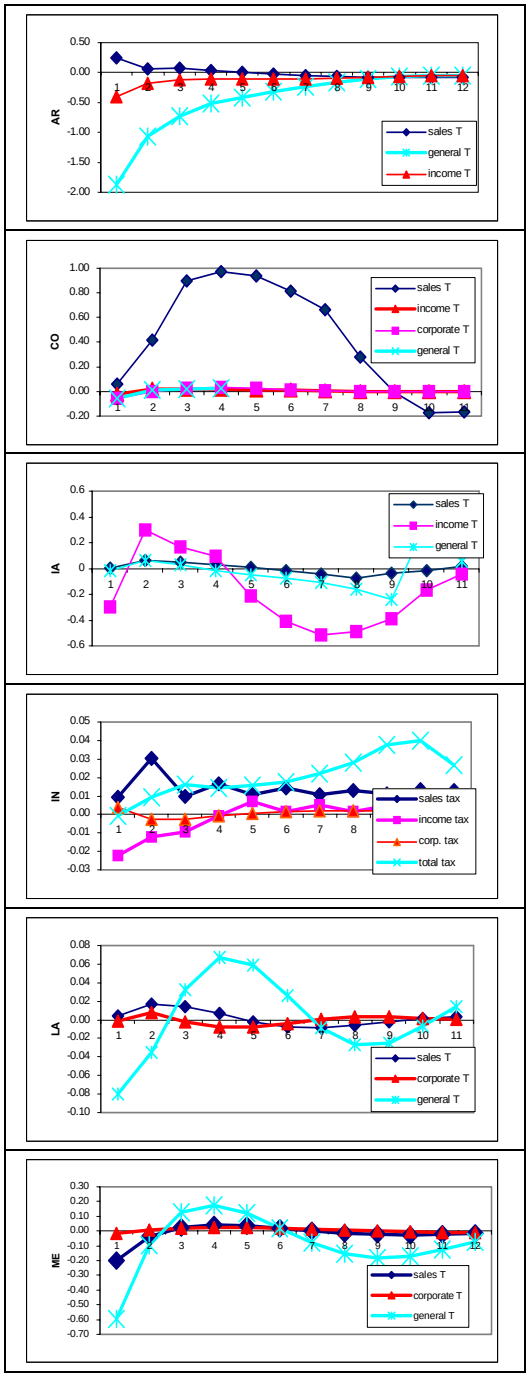


Figure A.8: Responses to Various Taxes

STATE	Size (area)	Size (Y/Y^{US})	G/Y	DEBT/Y	DF/USDF	ACIR index	Moody rates
AL	27	45	0.10	5.02	0.11	10	Aa3
AR	26	47	0.10	2.64	0.12	9	Aa2
AZ	5	32	0.09	1.73	0.14	10	A1
CA	2	7	0.09	3.88	0.17	6	Aa2
CO	7	13	0.07	2.07	0.11	10	A1
CT	46	6	0.08	11.25	0.17	5	Aa2
DE	47	3	0.10	14.25	0.14	10	Aaa
FL	25	35	0.07	3.32	0.11	10	Aa2
GA	20	25	0.08	3.06	0.13	10	Aaa
IA	22	26	0.09	1.86	0.13	10	Aaa
ID	10	41	0.10	3.31	0.12	10	Aaa
IL	23	9	0.07	4.39	0.14	4	Aa2
IN	37	27	0.08	2.47	0.13	10	Aa2
KS	12	19	0.08	1.49	0.12	10	Aa2
KY	35	38	0.10	8.53	0.12	10	Aaa
LA	32	12	0.09	7.82	0.08	4	A2
MA	44	10	0.10	10.02	0.18	3	Aa2
MO	17	22	0.07	3.16	0.11	10	Aaa
ME	38	43	0.11	8.34	0.13	9	Aa2
MI	21	17	0.09	3.53	0.15	6	Aaa
MN	13	15	0.10	3.78	0.19	8	Aaa
MD	41	1	0.09	7.08	0.16	6	Aaa
MS	30	48	0.11	4.69	0.11	9	Aa3
MT	3	37	0.11	6.13	0.07	10	Aa3
NB	14	20	0.08	2.39	0.12	10	Aa3
NC	28	30	0.09	2.22	0.14	10	Aaa
ND	16	33	0.12	4.20	0.10	8	A1
NM	4	28	0.12	4.50	0.13	10	Aa1
NJ	45	8	0.08	7.68	0.14	10	Aa2
NH	43	29	0.08	10.09	0.05	2	Aa2
NY	29	5	0.10	9.77	0.18	3	A2
NV	6	4	0.07	3.52	0.14	4	A1
OH	34	18	0.08	4.00	0.11	10	Aa1
OK	18	34	0.09	5.73	0.09	10	Aa3
OR	9	21	0.10	12.09	0.10	8	Aa2
PA	31	23	0.09	5.25	0.13	6	Aa2
RI	48	31	0.11	14.56	0.14	10	Aa2
SC	39	44	0.10	6.37	0.13	10	Aaa
SD	15	42	0.10	7.98	0.07	10	A1
TN	33	36	0.08	2.90	0.09	10	Aa2
TX	1	11	0.06	1.81	0.08	8	Aa1
UT	11	39	0.11	3.88	0.12	10	Aaa
VA	36	16	0.08	3.04	0.12	8	Aaa
VT	42	40	0.13	12.77	0.12	0	Aa1
WA	19	14	0.10	4.60	0.14	8	Aa2
WI	24	24	0.10	4.86	0.17	6	Aa3
WV	40	46	0.12	8.70	0.14	10	Aa3
WY	8	2	0.10	4.26	0.06	8	B
average			0.09	5.65	0.12		

Table A.1: Characteristics of US States

Country	Size (population)	Size (Y/Y^{EU})	G^c/Y	DEBT/Y	Country risk
Austria	8	7	0.19	66.3	A1
Belgium	6	6	0.22	113.9	A1
Finland	9	8	0.22	46.7	A1
France	2	2	0.24	58.4	A2
Germany	1	1	0.19	60.6	A2
Ireland	10	10	0.15	45.5	A1
Italy	3	3	0.18	113.0	A2
Netherlands	5	5	0.23	60.2	A2
Portugal	7	9	0.20	55.7	A2
Spain	4	4	0.18	60.9	A1
average			0.20	68.1	

Table A.2 Characteristics of EMU countries

Table TA2: Instantaneous US Price differential Responses and Yearly Multipliers

	G Shocks			BB shocks			T shocks		
	Price Y	Multiplier N	multiplier	Price Y	Multiplier N	multiplier	Price Y	Multiplier N	multiplier
AL	0.03	0.51	0.02				0.00	0.20	-0.001
AR	0.18	2.06	0.51				-0.40	1.21	0.03
AZ	-0.08	0.92	0.03				0.76	2.32	0.06
CA	0.09	1.71	0.02						
CO	0.01	1.62	0.06				-0.05	0.001	0.001
CT							0.25	0.72	-0.03
DE	0.01	2.91	0.01				0.03	0.45	0.005
FL	0.02	1.72	0.23						
GA	0.05	0.41	0.02				0.47	3.62	0.06
IA							-0.01	0.93	0.01
ID	-0.02	1.50	0.06				0.01	0.57	0.01
IL	-0.04	0.28	0.001				0.86	0.02	-0.001
IN	-0.03	2.14	0.03				-0.02	0.13	0.01
KS	-0.08	0.40	0.04	-0.12	-34.69	-0.26			
KY	0.06	0.75	0.12						
LA							-0.08	2.73	0.02
MA	-0.01	5.73	0.16						
MO				-0.34	-2.45	-0.09			
ME	-0.02	0.51	0.01				-0.60	6.57	0.05
MI	-0.01	2.09	0.02				0.60	1.41	-0.08
MN	-0.01	2.29	0.02						
MD	0.01	0.99	0.04	-0.18	-11.09	-0.27			
MS	-0.06	1.47	0.06	-0.26	-17.46	-0.001			
MT				-0.02	-1.50	0.01	0.07	0.21	-0.001
NB	0.06	6.54	0.06				0.04	0.43	0.01
NC	0.001	2.54	0.07				0.14	3.11	0.13
ND	0.001	4.72	-0.11				0.03	0.61	0.001
NM	0.08	23.60	0.13				0.01	0.10	0.001
NJ				-0.58	-3.78	0.17			
NH	0.01	0.26	0.07				0.11	0.27	-0.02
NY	-0.04	3.89	0.06						
NV	-0.05	1.03	-0.001				0.17	3.00	0.08
OH	0.00	1.20	0.07				-0.06	0.62	0.001
OK							0.22	9.06	0.09
OR	0.00	1.01	0.03				0.01	0.54	0.01
PA				-0.02	-0.10	0.01			
SC	0.01	0.98	0.03	-0.22	-10.56	-0.30			
SD	-0.03	10.80	0.08				0.06	2.06	0.001
TN							0.86	11.10	0.20
TX	0.05	1.40	0.02				0.19	7.96	0.12
UT	-0.03	0.70	0.03	-0.27	-14.44	-0.39			
VA							0.15	1.85	0.04
VT	-0.05	1.18	0.02						
WA	0.03	1.27	0.02	-0.09	-11.99	-0.15	0.59	7.91	0.27
WI	-0.02	1.25	0.01				0.05	0.10	0.01
WV	0.61	22.58	0.28	-0.04	-1.82	-0.001			
WY				-0.04	-0.80	0.05	0.05	0.64	-0.02
average	0.02	3.19	0.05	-0.18	-0.75	-0.28	0.14	2.27	0.03

Table TA3: Instantaneous EMU Price differential Responses and Yearly Multipliers

	G Shocks			BB shocks			T shocks		
	Price	Y Multiplier	N multiplier	Price	Y Multiplier	N multiplier	Price	Y Multiplier	N multiplier
Austria	0.55	2.41	0.19						
Belgium	3.66	6.30	-0.28						
Finland	0.79	2.53	0.97						
France							0.09	1.68	-0.01
Germany	-0.09	0.25	0.01				0.06	0.13	0.001
Ireland	-0.02	0.79	-0.07				0.25	9.20	0.06
Italy	0.01	1.35	-0.03				0.28	5.38	0.22
Netherland	0.07	0.62	0.01						
Spain	-0.08	4.69	0.04				0.10	0.60	0.01
Average	0.61	3.81	0.10				0.10	3.40	0.05

Bibliography

[1] Abdelkhalek, T. and Dufour, J.M. (1998), "Statistical Inference for CGE models with application to a model of the Moroccan economy", *Review of Economics and Statistics*, LXXX, 520-534.

[2] Ahmed, S., Ickes, B., Wang, P. and Yoo, B. (1993), "International Business Cycles", *American Economic Review*, 83, 335-359.

[3] Albert, J. and Chib, S. (1993), "Bayes Inference via Gibbs Sampling of Autoregressive Time Series Subject to Markov Mean and Variance Shifts", *Journal of Business and Economic Statistics*, 11, 1-16.

[4] Alesina, A. and Perotti, R. (1995), "Fiscal Expansions and Adjustments in OECD countries", *Economic Policy*, 21, 207-248.

[5] Altug, S. (1989), "Time to build and Aggregate Fluctuations: Some New Evidence", *International Economic Review*, 30, 883-920.

[6] Amemiya, T. (1985), *Advanced Econometrics*, Harvard University Press, Cambridge, Ma.

[7] Angeloni, I. Kashyap, A., Mojon, B. and Terlizzese, D. (2003), Monetary Transmission in the Euro area: where do we stand, in Angeloni, I. Kashyap, A., Mojon, B., Terlizzese, D. (eds.) *Monetary Transmission in the Euro area*, Cambridge University Press, Cambridge, UK.

[8] Anderson, B. and Moore, J. (1979), *Optimal Filtering*, Prentice Hall, Engelwood Cliffs, Ca.

[9] Andersson, T.W. (1971), *The Statistical Analysis of Time Series*, John Wiley & Sons, Chichister, UK.

[10] Anderson, T.W. and Hsiao, C. (1982), "Formulation and Estimation of Dynamic Models using Panel Data", *Journal of Econometrics*, 18, 47-82.

[11] Anderson, T. and Sörenson, B. (1996), "GMM estimation of Stochastic Volatility Model: A Monte Carlo Study", *Journal of Business and Economic Statistics*, 14, 328-352.

[12] Andrews, D., (1991), "Heteroskedasticity and Autocorrelation Consistent Covariance Matrix Estimation", *Econometrica*, 59, 817-858.

[13] Andrews, D. and Mohanan, C., (1992), "An Improved Heteroskedasticity and Autocorrelation Consistent Covariance Matrix Estimator", *Econometrica*, 60, 953-966.

[14] Arellano, M. and Bond, S. (1991), " Some tests of Specification for Panel data: Monte Carlo evidence and an Application to Employment Equations", *Review of Economic Studies*, 58, 277-297.

[15] Backus, D., Kehoe, P. and Kydland, F. (1994), "Dynamics of the Trade Balance and the Terms of Trade: The J-curve? *American Economic Review*, 84, 84-103.

[16] Backus, D., Kehoe, P. and Kydland, F. (1995), "International Business Cycles: Theory and Evidence", in T. Cooley (ed.), *Frontiers of Business Cycle Analysis*, Princeton University Press, Princeton, NJ, 331-356.

[17] Bai, J. (1997), "Estimation of multiple breaks one at a time", *Econometric Theory*, 13, 315-352.

[18] Ballabriga, F. (1997), "Bayesian Vector Autoregressions", ESADE manuscript.

[19] Baltagi, B. (1995), *Econometric Analysis of Panel Data*, Wiley & Sons, Chichester, UK.

[20] Barro, R., Mankiw, N and Sala, X. (1995), "Capital Mobility in Neoclassical Models of Growth", *American Economic Review*, 85, 103-115.

[21] Barro, R. and Sala, X. (1992), "Convergence", *Journal of Political Economy*, 100, 223-251.

[22] Bartle, R, (1976), *Elements of Real Analysis*, Wiley & Sons, Chichester, UK.

[23] Basu, S. and Fernand, J. (1997), " Returns to scale in US Production: Estimates and Implications" , *Journal of Political Economy* 105, 249-283.

[24] Baxter, M. and Crucini, M. (1993), "Explaining Saving-Investment Correlations", *American Economic Review*, 83, 416-436.

[25] Baxter, M. and King, R., (1999), "Measuring Business Cycles: Approximate Band-Pass Filters for Economic Time Series", *Review of Economics and Statistics*, 81, 575-593.

[26] Bayraktar, N., Sakellaris, P. and Vermeulen, P. (2003), "Real vs. Financial Frictions to Capital Investment", AUEB Athens, manuscript.

[27] Bauwens, L., M. Lubrano and J.F. Richard (1999) *Bayesian Inference in Dynamics Econometric Models*, Oxford University Press, Oxford, UK.

[28] Beaudry, P. and Portier, F. (2002), "The French Depression in the 1930's", *Review of Economic Dynamics*, 5, 73-99.

[29] Bell, W. (1984), "Signal Extraction for Nonstationary Time Series", *Annals of Statistics*, 12, 644-664.

[30] Benassy, J. P. (1995), "Money and Wage Contracts in a Optimizing Model of the Business Cycle", *Journal of Monetary Economics*, 35, 303-315.

[31] Benhabib, J. and Farmer, R. (2000), "The Monetary Transmission Mechanism", *Review of Economic Dynamics*, 3, 523-550.

[32] Berger, J. (1985), *Statistical Decision Theory and Bayesian Analysis*, Springer and Verlag, New York, NY.

[33] Berger, J. and Wolpert, R. (1998), *The Likelihood Principle*, Institute of Mathematical Statistics, Hayward, Ca., 2nd edition.

[34] Bernanke, B. and Boivin, J. and Eliasz, P. (2003), "Measuring the effects of Monetary Policy: A Factor Augmented VAR approach", NBER working paper 10220.

[35] Beveridge, S. and Nelson, C. (1981), "A New Approach to Decomposition of Economic Time Series into Permanent and Transitory Components with Particular Attention to the Measurement of the Business Cycle", *Journal of Monetary Economics*, 7, 151-174.

[36] Binder, M., Hsiao, C. and Pesaran, H. (2001), "Estimation and Inference in Short Panel VAR with Unit Roots and Cointergration", University of Maryland, manuscript.

[37] Blanchard, O. and Kahn, C. (1980), "The Solution of Difference Equations under Rational Expectations", *Econometrica*, 48, 1305-1311.

[38] Blanchard, O. and Quah, D. (1989), "The Dynamic Effect of Aggregate Demand and Supply Disturbances", *American Economic Review*, 79, 655-673.

[39] Boldrin, M. and Canova, F. (2001), "Inequality and Convergence: Reconsidering European Regional Policies", *Economic Policy*, 32, 205-253.

[40] Boldrin, M., Christiano, L. and Fisher, J. (2001), " Asset Pricing Lessons for Modelling Business Cycles", *American Economic Review*, 91, 146-166.

[41] Box, G. and Tiao, G. (1973), *Bayesian Inference in Statistical Analysis*, Wiley & Sons, Chichister, UK.

[42] Braun, P. and Mittnik, S. (1993), "Misspecifications in VAR and their effects on impulse responses and variance decompositions", *Journal of Econometrics*, 59, 319-341.

[43] Brockwell, P. and Davis, R. (1991), *Time Series: Theory and Methods*, Springer Verlag, New York, NY, second edition.

[44] Bry, G. and Boschen, C. (1971), *Cyclical analysis of time series: Selected Procedures and Computer Programs*, NBER, New York, NY.

[45] Burns, A. and Mitchell, W. (1946), *Measuring Business Cycles*, NBER, New York, NY.

[46] Burnside, C, Eichenbaum, M and Evans, C (1993), "Labor Hoarding and the Business Cycle", *Journal of Political Economy*, 101, 245-273.

[47] Burnside, C. and Eichenbaum, M. (1996), "Small Sample properties of GMM-based Wald Tests", *Journal of Business and Economic Statistics*, 14, 294-308.

[48] Campbell, J. (1994), "Inspecting the mechanism: An analytic approach to the stochastic growth model ", *Journal of Monetary Economics*, 33, 463-506.

[49] Campbell, J., Lo, A. and McKinley, C. (1997), *The Econonetrics of Finanicial Markets*, Princeton University Press, Princeton, NJ.

[50] Canova, F. (1991), Source of Financial Crisis: Pre and Post-Fed Evidence, *International Economic Review*, 32, 689-713.

[51] Canova, F. (1992), " An Alternative Approach to Modelling and Forecasting Seasonal Time Series " *Journal of Business and Economic Statistics*, 10, 97-108.

[52] Canova, F. (1993a), " Forecasting Time Series with Common Seasonal patterns", *Journal of Econometrics*, 55, 173-200.

[53] Canova, F. (1993b), "Forecasting exchange rates with a Bayesian time-varying coefficient model", *Journal of Economic Dynamics and Control*, 17, 233-261. .

[54] Canova, F. (1994), " Statistical Inference in Calibrated Models", *Journal of Applied Econometrics*, 9, S123-S144.

[55] Canova, F. (1995), "Sensitivity Analysis and Model Evaluation in Simulated Dynamic General Equilibrium Economies", *International Economic Review*, 36, 477-501.

[56] Canova, F. (1995), "VAR Models: Specification, Estimation, Inference and Forecasting", in H. Pesaran and M. Wickens (eds.), *Handbook of Applied Econometrics*, Ch.2, Blackwell, Oxford, UK.

[57] Canova, F. (1995), "The Economics of VAR Models", in K. Hoover (ed.), *Macroeconometrics: Tensions and Prospects*, Kluwer Press, NY, NY, 30-69.

[58] Canova, F. (1998), "Detrending and Business Cycle Facts", *Journal of Monetary Economics*, 41, 475-540.

[59] Canova, F. (1999), "Reference Cycle and Turning Points: A Sensitivity Analysis to Detrending and Dating Rules", *Economic Journal*, 109, 126-150.

[60] Canova, F. (2001), "Validating two DSGE monetary models with VARs", CEPR working paper 3442.

[61] Canova, F. (2002), G-7 Inflation forecasts, ECB working paper, 151.

[62] Canova, F. (2004), Testing for Convergence Club: A Predictive Density Approach, *International Economic Review*, 45, 49-77.

[63] Canova, F. and Ciccarelli, M., (2004), "Forecasting and Turning Point Prediction in a Bayesian Panel VAR model", *Journal of Econometrics*, 120, 327-359.

[64] Canova, F. and Ciccarelli, M., (2003), "Bayesian Panl VARs: Specification, Estimation, Testing and Leading Indicators", CEPR working paper 4033

[65] Canova, F. and De Nicolo, G. (2002), " Money Matters for Business Cycle Fluctuations in the G7", *Journal of Monetary Economics*, 49, 1131-1159.

[66] Canova, F. and De Nicoló, G. (2003), "The Equity Premium and the Risk Free Rate: A Cross Country, Cross Maturity Examination", *IMF Staff Papers*, 50, 250-285.

[67] Canova, F., Finn, M. and Pagan, A. (1994), "Evaluating a Real Business Cycle Model", in C. Hargreaves (ed.), *Nonstationary Time Series Analysis and Cointegration*, Oxford University Press, Oxford, UK, 225-255.

[68] Canova, F. and Marrinan, J. (1993), "Profits, Risk and Uncertainty in Exchange rates", *Journal of Monetary Economics*, 32, 259-286.

[69] Canova, F. and Marrinan, J. (1996), "Reconciling the term structure of Interest Rates with a consumption based I-CAP model ", *Journal of Economic Dynamics and Control*, 32, 259-286.

[70] Canova, F. and Ortega, E. (2000), "Testing Calibrated General Equilibrium Models", in Mariano, R., T. Shuermann and M. Weeks (eds.) *Inference using Simulation Techniques*, Cambridge University Press, Cambridge, UK, 400-436.

[71] Canova, F. and Pappa, E. (2003), "Price Differential in Monetary Unions: the Role of Fiscal shocks", CEPR working paper 3746.

[72] Canova, F. and Pina, J. (2004), "What VARs tell us about DSGE models?", forthcoming in C. Diebolt, and Krystou, C. (eds.), *New Trends In Macroeconomics*, Springer Verlag, New York, NY.

[73] Carlin B.P. and Gelfand, A.E, Smith, A.F.M (1992) "Hierarchical Bayesian Analysis of change point problem", *Journal of the Royal Statistical Society*, Ser C, 389-405.

[74] Carlin, B., Polsom, N. and Stoffer, D. (1992), "A Monte Carlo Approach to nonnormal and nonlinear state-space modelling", *Journal of the American Statistical Association*, 87, 493-500.

[75] Carter, C. and Kohn, P. (1994) "On the Gibbs sampling for State Space Models", *Biometrika*, 81, 541-553.

[76] Casella, G. and George, E. (1992), "Explaining the Gibbs Sampler" *American Statistician*, 46, 167-174.

[77] Cecchetti, S.G., Lam, P. and Mark, N. (1993), "The Equity Premium and the Risk Free Rate: Matching Moments", *Journal of Monetary Economics*, 31, 21-45.

[78] Chari, V., Kehoe, P. and McGrattan, E. (2000) "Sticky Price Models of the Business Cycle: Can the Contract Multiplier Solve the Persistence Problem?", *Econometrica*, 68, 1151-1179.

[79] Chib, S. (1995) "Marginal Likelihood from the Gibbs Output ", *Journal of the American Statistical Association*, 90, 1313-1321.

[80] Chib, S. (1996) "Calculating Posterior Distributions and Model Estimates in Markov Mixture Models", *Journal of Econometrics*, 75, 79-98.

[81] Chib, S. and Greenberg, E. (1995), "Understanding the Hasting-Metropolis Algorithm", *The American Statistician*, 49, 327-335.

[82] Chib, S. and Greenberg, E. (1996), "Markov chain Monte Carlo Simulation methods in Econometrics", *Econometric Theory*, 12, 409-431.

[83] Cho, J. and Cooley, T. (1995), "The Business Cycle with Nominal Contracts", *Economic Theory*, 6, 13-33.

[84] Christiano, L. and Eichenbaum, M. (1992), "Current Real Business Cycle Theories and Aggregate Labor Market Fluctuations", *American Economic Review*, 82, 430-450.

[85] Christiano, L., Eichenbaum, M. and Evans, C. (2001), "Nominal Rigidities and the Dynamic Effects of a shock to Monetary Policy", forthcoming, *Journal of Political Economy*.

[86] Christiano, L. and Fitzgerard, T. (2003), "The Band Pass Filter", *International Economic Review*.

[87] Christiano, L., Gust, C. and Roldos, J. (2001), "Monetary Policy in a Financial Crisis", forthcoming, *Journal of Economic Theory*.

[88] Christiano, L. and den Haan, W. (1996), "Small Sample Properties of GMM for Business Cycle Analysis", *Journal of Business and Economic Statistics*, 14, 309-327.

[89] Ciccarelli, M. and Rebucci, A. (2002), "Has the transmission Mechanism of Monetary Policy Changed over time", forthcoming, *European Economic Review*.

[90] Ciccarelli, M. and Rebucci, A. (2003), " Measuring Contagion with a Bayesian TVC model", ECB working paper 263.

[91] Clark, P. (1973), "Subordinated Stochastic Process Model with Finite Variance for Speculative Prices", *Econometrica*, 41, 136-156.

[92] Cogley, T. and Nason, J. (1994), "Testing the Implications of Long-run Neutrality for Monetary Business Cycle Models", *Journal of Applied Econometrics*, 9, S37-S70.

[93] Cogley, T. and Nason, J., (1995), "The Effects of the Hodrick and Prescott Filter on Integrated Time Series", *Journal of Economic Dynamics and Control*, 19, 253-278.

[94] Cogley, T. and Nason, J.M. (1995), "Output Dynamics in Real Business Cycle Models", *American Economic Review*, 85, 492-511.

[95] Cogley, T. and Sargent, T. (2002), "Bayesian Prediction Intervals in Evolving Monetary Systems", NYU university, manuscript.

[96] Cooley, T. (1995), (ed.) *Frontiers of Business Cycle Research*, Princeton University Press, Princeton, NJ.

[97] Cooley, T. and Dwyer, M. (1998), "Business Cycle Analysis without much Theory: A Look at Structural VARs", *Journal of Econometrics*, 83, 57-88.

[98] Cooley, T. and Hansen, G. (1989), "The Inflation Tax in a Real Business Cycle Model", *American Economic Review* 79, 733-748.

[99] Cooley, T. and LeRoy, S. (1985), "Atheoretical macroeconomics: A critique", *Journal of Monetary Economics*, 16, 283-308.

[100] Cooley, T. and Prescott, E. (1973), "Estimation in the Presence of Stochastic Parameter Variation", *Econometrica*, 44, 167-184.

[101] Corbae, D. and Ouliaris, S. (2001), "Extracting Cycles from Nonstationary Data", IMF manuscript.

[102] Corbae, D., Ouliaris, S., and Phillips, P. (2002), "Band Spectral Regression with Trending Data", *Econometrica*, 70, 1067-1109.

[103] Cumby, R., Obstfeld, M and Huizinga, J. (1982), "Two step, Two stage least square estimation in models with rational expectations", *Journal of Econometrics*, 21, 333-353.

[104] Danthine, J.P. and Donaldson, J. (1992), "Non-Walrasian Economies", Cahiers de Recherche Economique, Université de Lausanne, No.9301.

[105] Davidson, J. (1994), *Stochastic Limit Theory*, Oxford University Press, Oxford, UK.

[106] Dedola, L. and Neri, S. (2003), " What are the effects of technology shocks? A VAR analysis with prior on impulse responses", Bank of Italy and ECB, manuscript.

[107] DeJong, D., Ingram, B. and Whiteman, C., (1996), "Beyond Calibration", *Journal of Business and Economic Statistics*, 14, 1-10.

[108] DeJong, D. Ingram, B. and Whiteman, C. (2000), "A Bayesian Approach to Dynamic Macroeconomics ", *Journal of Econometrics*, 98, 203-223.

[109] Del Negro, M. and Schorfheide, F. (2004), " Priors from General Equilibrium Models for VARs", *International Economic Review*.

[110] Den Haan, W. and Levine, A. (1996), " Inference from parametric and nonparametric covariance matrix estimation procedures", UCSD manuscript.

[111] Diebold, F. and Mariano, R. (1995), "Predictive Accuracy", *Journal of Business and Economic Statistics*, 13, 253-265.

[112] Diebold, F., Ohanian, L. and Berkowitz, J., (1998), "Dynamics General Equilibrium Economies: A Framework for Comparing Models and Data", *Review of Economic Studies*, 68, 433-451.

[113] Doan, T., Litterman, R and Sims, C. (1984), "Forecasting and Conditional Projection Using Realistic Prior Distributions", *Econometric Reviews*, 3, 1-100.

[114] Doan, T. (2000), Rats 5.0 Manual, Estima, Il.

[115] Dotsey, M., King, R. and Wolmann, A. (1999), "State Dependent Pricing and the General Equilibrium Dynamics of Money and Output, *Quarterly Journal of Economics*, 114, 655-690.

[116] Dridi, R. and Renault, E (1998), "Semiparametric Indirect Inference", University of Toulouse, manuscript.

[117] Dridi, R. , Guay, A. and Renault, E. (2003), "Indirect Inference and Calibration of Dynamic Stochastic General Equilibrium Models", University of Montreal, manuscript.

[118] Duffie, D. and Singleton, K. (1993), "Simulated Moments Estimation of Markov Models of Asset Prices", *Econometrica*, 61, 929-950.

[119] Edge, R. (2002), "The equivalence of Wage and Price Staggering In Monetary Business Cycle Models", *Review of Economic Dynamics*, 5, 559-585.

[120] Eichenbaum, M., Hansen, L. and Singleton, K. (1988), " A Time Series analysis of the representative Agent Models of Consumption and Leisure choice under uncertainty", *Quarterly Journal of Economics*, 103, 51-78.

[121] Engle, R. (1974), "Band Spectrum Regression", *International Economic Review*, 15, 1-11.

[122] Engle, R. (1983), "Wald, Likelihood ratio and Lagrange Multiplier Tests in Econometrics" in Z. Griliches, M. Intrilligator (eds.) *Handbook of Econometrics*, vol II, Amsterdaam, North Holland, 775-826.

[123] Engle, R., Lilien, D. and Robins, R. (1987), Estimating Time Varying Risk Premia in Term Structures: The ARCH-M model, *Econometrica*, 55, 391-408.

[124] Erceg, C., Henderson, D. and Levin, A. (2000), "Optimal Monetary Policy with Staggered wage and price contracts", *Journal of Monetary Economics*, 46, 281-313.

[125] Evans, G. and Reichlin, L., (1994), "Information, Forecasts and The Measurement of the Business Cycles", *Journal of Monetary Economics*, 33, 233-254.

[126] Farmer, R.(1997) "Money in a RBC model", *Journal of Money Banking and Credit*, 29, 568-611.

[127] Fatas A. and Mihov, I. (2001), "Government Size and the Automatic Stabilizers: International and Intranational Evidence," *Journal of International Economics*, 55, 2-38.

[128] Favero, C. (2001), *Applied Macroeconometrics* , Oxford University Press, Oxford, UK.

[129] Faust, J. (1998), " On the Robustness of Identified VAR Conclusions about Money" , *Carnegie-Rochester Conference Series on Public Policy*, 49, 207-244.

[130] Faust, J. and Leeper, E. (1997), "Do Long Run Restrictions Really Identify Anything?", *Journal of Business and Economic Statistics*, 15, 345-353.

[131] Fernandez Villaverde, J. and Rubio Ramirez, J. (2003a), "Estimating Dynamic Equilibrium Economies: A Likelihood Approach", Federal Reserve of Atlanta, manuscript.

[132] Fernandez Villaverde, J. and Rubio Ramirez, J. (2003b), "Estimating Dynamic Equilibrium Economies: Linear vs. Nonlinear Likelihood", Federal Reserve of Atlanta, manuscript

[133] Fernandez-Villaverde, J. and Rubio-Ramirez, J. (2004), "Comparing Dynamic Equilibrium Models to the data", *Journal of Econometrics*, 123, 153-187.

[134] Ferson, W. and Foerster, S. (1994), "Finite Sample Properties of GMM in Tests of Conditional Asset Pricing Models", *Journal of Financial Economics*, 36, 29-55.

[135] Fève, P. and Langot, F. (1994), "The RBC Models through Statistical Inference: An Application with French Data", *Journal of Applied Econometrics*, 9, S11-S37.

[136] Finn, M. (1998), "Cyclical Effects of Goverment's Employment and Goods Purchases", *International Economic Review*, 39, 635-657.

[137] Fruhwirth-Schnatter, S. (2001), "MCMC estimation of classical and Dynamic switching and Mixture Models", *Journal of the American Statistical Association*, 96, 194-209.

[138] Furher, J. Moore, J. and Schuh, S. (1995), "Estimating the Linear Quadratic Inventory model, ML vs GMM", *Journal of Monetary Economics*, 35, 115-157.

[139] Gali, J., (1992), "How Well Does the ISLM Model Fit Postwar U.S. Data?", *Quarterly Journal of Economics*, CVII, 709-738.

[140] Gali, J. (1999), Technology, Employment and Business Cycle: Do Technology Shocks Explain Aggregate Fluctuations?, *American Economic Review*, 89, 249-271.

[141] Gali, J and Gertler, M. (1999), "Inflation Dynamics: A Structural Econometric Analysis, *Journal of Monetary Economics*, 44, 195-222.

[142] Gallant, R. (1987), *Nonlinear Statistical Models*, Wiley & Sons, Chichister, UK.

[143] Gallant, R. and Tauchen, G., (1996), "Which Moments to Match?", *Econometric Theory*, 12, 657-681.

[144] Gallant, A.R., Rossi P. and Tauchen G. (1993), "Nonlinear Dynamic Structures, *Econometrica*, 61, 871-908.

[145] Gelfand, A. and Dey, D. (1990), "Bayes Model Choice: Asymptotics and Exact Calculations", *Journal of the Royal Statistical Society*, Ser B, 501-514.

[146] Gelfand, A. and Smith, A. F. M. (1990), "Sampling-Based Approaches to Calculating Marginal Densities", *Journal of The American Statistical Association*, 85, 398-409.

[147] Gelman, A., J. B. Carlin, H.S. Stern and D.B. Rubin (1995), *Bayesian Data Analysis*, Chapman and Hall, London, UK.

[148] Gertler, M. and Girlchrist, S. (1994), Monetary Policy, Business Cycles and the Behavior of Small Manufactoring Firms, *Quarterly Journal of Economics*, CIX, 309-340.

[149] Geweke, J. (1989), "Bayesian Inference in Econometric Models Using Monte Carlo Integration", *Econometrica*, 57, 1317-1339.

[150] Geweke, J. (1994), Comment to Jacquier, Polson and Rossi , *Journal of Business and Economic Statistics*, 12, 397-398.

[151] Geweke, J. (1995), " Monte Carlo Simulation and Numerical Integration" in Amman, H., Kendrick, D. and Rust, J. (eds.) *Handbook of Computational Economics*, Elsevier, St. Louis, Mo., 731-800.

[152] Geweke, J. (1999), "Computational Experiment and Reality", University of Iowa, manuscript.

[153] Geweke, J. and Zhou, G. (1996), "Measuring the Pricing Error of the Arbitrage Pricing Theory", *Review of Financial Studies*, 9, 557-587.

[154] Giannone, D., Reichlin, L and Sala, L. (2003), "Tracking Greenspan: Systematic and Unsystematic Monetary Policy revisited", CEPR working paper 3550.

[155] Giordani, P. (2002), "An Alternative Explanation of the Price Puzzle", forthcoming, *Journal of Monetary Economics*.

[156] Gomez, V. (1997), "Three equivalent methods for Filtering Nonstationary Time Series", *Journal of Business and Economic Statistics*, 17, 109-166.

[157] Gordin, M. (1969), "The Central Limit Theorem for Stationary Processes," *Soviet Math. Doke.,* pp. 1174-1176.

[158] Gordon, D. and Leeper, E. (1994), "The dynamic Impact of Monetary Policy: An Exercise in Tentative Identification", *Journal of Political Economy*, 102, 1228-1247.

[159] Gourieroux, C. and Monfort, A. (1995), "Testing, Encompassing and Simulating Dynamic Econometric Models", *Econometric Theory*, 11, 195-228.

[160] Gourieroux, C., Monfort, A. and Renault, E. (1993), "Indirect Inference", *Journal of Applied Econometrics*, 8, S85-S118.

[161] Gourinchas, J. and Jeanne, O. (2003), "The Elusive Gains from International Financial Integration", IMF manuscript.

[162] Granger, C. and Anderson, A. (1978), *An Introduction to Bilinear Time Series Models*, Vandenhoeck and Ruprecth, Gottingen, Sweden

[163] Gregory, A. and Smith, G. (1989), "Calibration as Estimation", *Econometric Reviews*, 9(1), 57-89.

[164] Gregory, A. and Smith, G. (1991), "Calibration as Testing: Inference in Simulated Macro Models", *Journal of Business and Economic Statistics*, 9(3), 293-303.

[165] Gregory, A. and Smith, G. (1993), "Calibration in Macroeconomics", in Maddala, G.S. (ed.), *Handbook of Statistics*, vol. 11, Elsevier, St.Louis, Mo., 703-719.

[166] Hall, A., (1992), "Some Aspects of Generalized Method of Moment Estimators", in G.S. Maddala, C.R. Rao and H.D. Vinod (eds.), *Handbook of Statistics*, vol. 11, Elsevier, St. Louis, Mo., 653-685.

[167] Hamilton, J. (1994) *Time Series Analysis*, Princeton University Press, Princeton, N.J.

[168] Hamilton, J. (1989) "A New Approach to the economic analysis of nonstationary time series and the business cycle", *Econometrica*, 57, 357-384.

[169] Hansen, L.P., (1982), "Large Sample Properties of GMM Estimators", *Econometrica*, 50, 1029-1054.

[170] Hansen, L.P., (1985), "A Method for Calculating Bounds on the Asymptotic Covariance Matrix of GMM Estimators", *Journal of Econometrics*, 30, 203-231.

[171] Hansen, L. and Heckman, J. (1996) "The Empirical Foundations of Calibration", *Journal of Economic Perspective*, 10, 87-104.

[172] Hansen, L., Heaton, J. and A. Yaron, (1996) " Finite Sample Properties of Alternative GMM Estimators", *Journal of Business and Economic Statistics*, 14, 262-281.

[173] Hansen, L. and Hodrick, R. (1980) "Forward Exchange Rates as Optimal Predictors of Future Spot Rates: An Econometric Analysis", *Journal of Political Economy*, 88, 829-853.

[174] Hansen, L. and Jagannathan, R. (1991), "Implications of Security Market Data for Models of Dynamic Economies", *Journal of Political Economy*, 99, 225-262.

[175] Hansen, L. and Sargent, T. (1979), "Formulating and Estimating Dynamic Linear Rational Expectations Models", *Journal of Economic Dynamic and Control*, 2, 7-46.

[176] Hansen, L. and Sargent, T. (1982), "Instrumental Variables Procedures for Linear Rational Expectations Models", *Journal of Monetary Economics*, 9, 263-296.

[177] Hansen, L. and Sargent, T., (1991), "Two Difficulties in Interpreting Vector Autoregressions", in Hansen, L. and Sargent, T. (eds.), *Rational Expectations Econometrics*, Westview Press: Boulder & London.

[178] Hansen, L. and Sargent, T. (1998), *Recursive Linear Models of Dynamic Economies*, Princeton University Press, Princeton, NJ.

[179] Hansen, L. , Sargent, T. and McGrattan, E. (1996), " Mechanics of Forming and Estimating Dynamic Linear Economies", in Amman, H., Kendrick, D. and Rust, J. *Handbook of Computational Economics*, Elsevier, St. Louis, Mo.

[180] Hansen, L. and Singleton, K., (1982), "Generalized Instrumental Variables Estimation of Nonlinear Rational Expectations Models", *Econometrica*, 50, 1269-1286. (corrigenda, 1984).

[181] Hansen, L.P. and Singleton, K., (1988), " Efficient Estimation of Linear Asset Pricing Models with Moving Averages Errors", manuscript.

[182] Harvey, A. (1985), "Trends and Cycles in Macroeconomic Time Series", *Journal of Business and Economic Statistics*, 3, 216-227.

[183] Harvey, A.C. (1991), *Time Series Models*, Philip Allan, Deddigton, UK.

[184] Harvey, A. and Jeager, A., (1993), "Detrending, Stylized Facts and the Business Cycles", *Journal of Applied Econometrics*, 8, 231-247.

[185] Hastings, N.A.J. and Peacock. J.B. (1980) *Statitical Distributions*, Wiley & Sons, New York, NY.

[186] Hausman, J., (1978), "Specification Tests in Econometrics", *Econometrica*, 46, 1251-1271.

[187] Hayashi, F. (2002), *Econometrics*, Princeton University Press, Princeton, NJ.

[188] Hayashi, F. and Sims, C. (1983), "Nearly Efficient Estimation in Time Series Models with Predetermined, but not Exogenous Instruments", *Econometrica*, 51, 783-798.

[189] Hess, G. and Iwata, S. (1997), " Measuring and Comparing Business Cycle Features", *Journal of Business and Economic Statistics*, 15, 432-444.

[190] Hodrick, R. and Prescott, E., (1997), "Post-War US Business Cycles: An Empirical Investigation", *Journal of Money Banking and Credit*, 29, 1-16.

[191] Holtz-Eakin, D. (1988), "Testing for Individual Effects in Autoregressive Models", *Journal of Econometrics*, 39, 297-307.

[192] Holtz Eakin, D., Newey, W. and Rosen, H. (1988), " Estimating Vector Autoregression with Panel data", *Econometrica*, 56, 1371-1395.

[193] Hsiao, C. (1989), *Analysis of Panel Data*, Cambridge University Press, Cambridge, U.K.

[194] Jacquier, E., Polson N. and Rossi, P. (1994), " Bayesian Analysis of Stochastic Volatility Models", *Journal of Business and Economic Statistics*, 12, 371-417.

[195] Jeffreys, H. (1966), *Theory of Probability* (3rd edition), Clarendon, Oxford, UK.

[196] Journal of Business and Economic Statistics, January 1990.

[197] Judge, G, Griffith, W., Carter Hill, R., Lee, and Lutkepohl, H.(1985), *Theory and Practice of Econometrics*, Wiley & Sons, Second Edition.

[198] Judson, R. and Owen, A. (1999), "Estimating Dynamic Panel Data Models: A Practical Guide for Macroeconomists", *Economic Letters*, 65, 145-150.

[199] Kadiyala, R. and Karlsson, S. (1997), "Numerical methods for estimation and Inference in Bayesian VAR models", *Journal of Applied Econometrics*, 12, 99-132.

[200] Kass, R. and Raftery, A (1995), "Empirical Bayes Factors", *Journal of the American Statistical Association*, 90, 773-795.

[201] Kauffman, S. (2003), "Business Cycle of European Countries. Bayesian clustering of country-individual IP growth series", Oesterreichische Nationalbank, manuscript.

[202] Keane, M. and Runkle, D. (1992) "On the estimation of Panel data Models with Serially Correlation when Instruments are not Strictly Exogenous, *Journal of Business and Economic Statistics*, 10, 1-9.

[203] Kilian, L. (1998), "Small Sample confidence Intervals for Impulse Response Functions", *Review of Economics and Statistics*, 218-230.

[204] Kilian, L. and Ivanov, V. (2001), "A Practitioner's Guide to Lag Order Selection for VAR Impulse Response Analysis", ECB working paper 195.

[205] Kim, C. and Nelson, C. (1999), *State Space Models with Regime Switching*, MIT Press, Candridge, Ma.

[206] Kim, J. (2000) "Constructing and Estimating a realistic Optimizing Model of Monetary Policy", *Journal of Monetary Economics*, 45, 329-359

[207] Kim, K. and Pagan, A. (1994) "The Econometric Analysis of Calibrated Macroeconomic Models", in Pesaran, H. and M. Wickens (eds.), *Handbook of Applied Econometrics*, Vol.I, Blackwell Press, Oxford, UK, 356-390.

[208] King, R. and Plosser, C. (1994) "Real Business Cycles and the Test of the Adelmans", *Journal of Monetary Economics*, 33, 405-438.

[209] King, R., Plosser, C. and Rebelo S. (1988), "Production, Growth and Business Cycles: I and II", *Journal of Monetary Economics*, 21, 195-232 and 309-342.

[210] King, R., Plosser, C. and Rebelo, S. (1988), Appendix to "Production, Growth and Business Cycle: I The Basic Neoclassical models", University of Rochester, working paper.

[211] King, R. Plosser, C., Stock, J. and Watson, M. (1991), "Stochastic Trends and Economic Fluctuations", *American Economic Review*, 81, 819-840.

[212] King, R. and Rebelo, S., (1993), "Low Frequency Filtering and Real Business Cycles", *Journal of Economic Dynamics and Control*, 17, 207-231.

[213] King, R. and Watson, M. (1998), "The solution of Singular Linear Difference Systems under Rational Expectations", *International Economic Review*, 39, 1015-1026.

[214] Kiviet, J. (1995), " On bias, Inconsistency and efficiency of various estimators in Dynamic Panel Data Models", *Journal of Econometrics*, 68, 53-78.

[215] Kiyotaki, N. and Moore, J. (1997), "Credit Cycles", *Journal of Political Economy*, 105, 211-248.

[216] Klein, P. (2001), " Using the generalized Schur form to solve a multivariate linear rational expectations model", *Journal of Economic Dynamics and Control*, 24, 1405-1423.

[217] Kocherlakota, N. (1990), "On Tests of Representative Consumer Asset Pricing Models", *Journal of Monetary Economics*, 26, 285-304.

[218] Koop, G. (1996), "Bayesian Impulse responses", *Journal of Econometrics*, 74, 119-147.

[219] Koop, G., Osiewalski, J. and Steel, M. (2000), "Modelling the Sources of Output Growth in a Panel of Countries", *Journal of Business and Economic Statistics*, 18, 284-299.

[220] Koop, G., Pesaran, H. and Potter, S. (1996), "Impulse Response Analysis in Nonlinear Multivariate Models", *Journal of Econometrics*, 74, 119-147.

[221] Kuhn, T. (1970), *The Structure of Scientific Revolutions*, Chicago University Press, Chicago, Il.

[222] Kurmann, A. (2003), "ML estimation of Dynamic Stochastic Theories with an Application to New Keynesian Pricing", University of Quebec at Montreal, manuscript.

[223] Kydland, F. and Prescott, E. (1982), "Time To Build and Aggregate Fluctuations", *Econometrica*, 50, 1345-1370 .

[224] Kydland, F. and Prescott, E. (1991), "The Econometrics of the General Equilibrium Approach to Business Cycles", *The Scandinavian Journal of Economics*, 93(2), 161-178.

[225] Kydland, F. and Prescott, E. (1996) "The Computational Experiment: An Econometric Tool", *Journal of Economic Perspective*, 10, 69-85.

[226] Imbs, J. (2002), "Why the link between Volatility and Growth is both Positive and Negative", LBS manuscript.

[227] Ingram, B. and Lee, B.S. (1989) "Estimation by Simulation of Time Series Models", *Journal of Econometrics*, 47, 197-207.

[228] Ingram, B. and Whitemann, C. (1994), "Supplanting the Minnesota prior. Forecasting macroeconomic time series using real business cycle priors", *Journal of Monetary Economics*, 34, 497-510.

[229] Ireland, P. (2000), "Sticky Price Models and the Business Cycle: Specification and Stability", *Journal of Monetary Economics*, 47, 3-18.

[230] Ireland, P. (2004), "A method for taking Models to the data", *Journal of Economic Dynamics and Control*, 28, 1205-1226.

[231] Lahiri, K. and Moore, G. (1991), *Leading Indicators: New Approaches and Forecasting Record*, University of Chicago Press, Chicago, Il.

[232] Lam, P. (1990), "The Hamilton model with general autoregressive Component", *Journal of Monetary Economics*, 26, 409-432.

[233] Leeper, E. (1991), "Equilibria Under Active and Passive Monetary and Fiscal Policies", *Journal of Monetary Economics*, 27, 129-147.

[234] Leeper, E. and Sims, C. (1994), " Towards a Modern Macroeconomic Model Usable for Policy Analysis", in Rotemberg, J. and Fisher, S. (eds.) *NBER Macroeconomic Annual*, 9, 81-118.

[235] Leeper, E. Sims, C. and Zha, T. (1996), " What does Monetary Policy Do", *Brookings Papers of Economic Activity*, 2, 1-78.

[236] Ljung, L. and Soderstroem, T. (1983), *Theory and Practice of Recursive Identification*, Cambridge, MIT Press, Ma.

[237] Linde', J. (2001), "Estimating New Keynesian Phillips curve: A Full Information Maximum Likelihood", Riksbank, working paper 129.

[238] Lindlay, D. V. and Smith, A.F.M. (1972) "Bayes Estimates of the Linear Model", *Journal of the Royal Statistical Association*, Ser B, 34, 1-18.

[239] Lippi, M. and Reichlin, L. (1993) "The Dynamic Effect of Aggregate Demand and Supply Disturbances: A Comment", *American Economic Review*, 83, 644-652.

[240] Lippi, M. and Reichlin, L., (1994), "VAR Analysis, Non-Fundamental Representation, Blaschke Matrices", *Journal of Econometrics*, 63, 307- 325.

[241] Loeve, M., (1974), *Probability Theory*, Springer Verlag, New York, NY.

[242] Lucas, R. (1980), "Two illustrations in the quantity theory of money", *American Economics Review*, 70, 1345-1370.

[243] Lucas, R. and Stockey, N. (1989), *Recursive Methods in Economic Dynamics*, Harvard University Press, Cambridge, Ma.

[244] Lukacs, E. (1975), *Stochastic Convergence*, Academic Press, Burlington, Ma.

[245] Lutkepohl, H., (1991), *Introduction to Multiple Time Series Analysis*, Second edition, Springer and Verlag, New York, NY.

[246] MacKinley, C. and Richardson, M. (1991), "Using GMM to test Mean-Variance Efficiency", *The Journal of Finance*, XLIV, 511-527.

[247] Maddala, G. S. and Hu, W. (1996), "The Pooling Problem" in L. Matyas and P. Sevestre (eds.) *Econometrics of Panel Data*, Kluwer, New York, NY.

[248] Maffezzoli, M. (2000) "Human Capital and International Business Cycles", *Review of Economic Dynamics*, 3, 137-165.

[249] Malinvaud, E. (1980), *Statistical Methods in Econometrics*, 3rd edition, North Holland, Amsterdaam, The Netherland.

[250] Mao, C. S. (1990) "Hypothesis testing and Finite Sample Properties of GMM Estimators: A Monte Carlo Study", Federal Reserve Bank of Richmond, manuscript.

[251] Maravall, A. and Del Rio, A. (2001) "Time aggregation and the Hodrick-Prescott Filter", Bank of Spain, manuscript.

[252] Marcet, A. (1991), "Time Aggregation of Econometric Time Series ", in Hansen, L. and Sargent, T., (eds.), *Rational Expectations Econometrics*, Westview Press: Boulder & London.

[253] Marcet, A. (1992), "Solving Nonlinear Stochastic Models by Parametrizing Expectations: An Application to Asset Pricing with Production", Universitat Pompeu Fabra, working paper 5.

[254] Marcet, A. and Den Haan, W (1994) "Accuracy in Simulation", *Review of Economic Studies*, 61, 3-17.

[255] Marcet, A. and Lorenzoni, G. (1999), "The parametrized expectations approach: Some Practical issues", in Marimon, R. and Scott, A. (eds.) *Computational Methods for the Study of Dynamic Economies*, Oxford University Press, Oxford, UK.

[256] Marcet, A. and Ravn, M. (2001) "The HP filter in Cross Country Comparisons", LBS manuscript.

[257] Marimon R. and Scott, A. (1999), (eds.) *Computational Methods for the Study of Dynamic Economies*, Oxford University Press, Oxford, UK.

[258] Martin, V. and Pagan, A. (2001), "Simulation Based estimation of some factor models in econometrics" in Mariano, R. Schuermann and M. Weeks (eds.) *Inference Using Simulation Techniques*, Cambridge University Press, Cambridge, UK.

[259] McCulloch, R. and Tsay, R. (1994), "Statistical Analysis of Economic Time Series via Markov Switching Models", *Journal of Time Series Analysis*, 15, 521-539.

[260] McGrattan, E. Rogerson, R. and Wright, R. (1987), "An Equilibrium Model of the Business Cycle with Household Production and Fiscal Policy", *International Economic Review*, 38, 267-290.

[261] McLeish, R. (1974) "Dependent Central Limit Theorem and Invariance Principle", *Annals of Probability*, 2, 620-628.

[262] Merha, R. and Prescott, E. (1985), "The Equity Premium: A Puzzle", *Journal of Monetary Economics*, 15, 145-161.

[263] Merz, M. (1995) "Search in Labor Markets and Real Business Cycles", *Journal of Monetary Economics*, 36, 269-300.

[264] Mittnik, S. and Zadrozky, P. (1993), "Asymptotic Distributions of Impulse Responses, Step Responses and Variance Decompositions of Estimated Linear Models", *Econometrica*, 61, 857-871.

[265] Morley, J., Nelson, C. and Zivot, E. (2003), " Why are Beveridge-Nelson and Unobservable Component Decompositions of GDP so Different? ", *Review of Economics and Statistics*, 86,

[266] Morris, C. (1983) "Parametric Empirical Bayes Inference: Theory and Applications", *Journal of the American Statistical Association*, 78, 47-59.

[267] Murray, C. (2002), "Cyclical Properties of Baxter and King filtered Time Series", *Review of Economics and Statistics*, 85, 472-476.

[268] Neiss, K. and Pappa, P. (2002) "Persistence without too much stickiness: The role of factor utilization", forthcoming, *Review of Economic Dynamics*.

[269] Nelson, C. and Kang, H. (1981), "Spurious Periodicity in appropriately detrended time series", *Econometrica*, 49, 741-751.

[270] Nelson, C. and Starz, R. (1990), "The Distribution of the Instrumental Variable Estimator and its t-ratio when the Instrument is a poor one", *Journal of Business*, 63, 125-164.

[271] Newey, W. (1990) "Efficient Instrumental variable estimation of nonlinear models", *Econometrica*, 58, 809-837.

[272] Newey, W. and West, K., (1987), "A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix", *Econometrica*, 55, 703-708.

[273] Newey, W. and West, K., (1994), "Automatic Lag Selection in Covariance Matrix Estimation", *Review of Economic Studies*, 61, 631-653.

[274] Newey, W. and McFadden, D. (1994), "Large Sample Estimation and Hypothesis Testing", in Engle, R., and McFadden, D. (eds.) *Handbook of Econometrics*, vol IV, Elsevier, St. Louis, Mo, 2111-2245.

[275] Nickell. S. (1981), "Biases in Dynamic Models with fixed Effects", *Econometrica*, 49, 1417-1426.

[276] Obsorn, D. (1995), "Moving Average Detrending and the Analysis of Business Cycles", *Oxford Bulletin of Economics and Statistics*, 57, 547-558.

[277] Ogaki, M., (1993), "GMM: Econometric Applications", in G.S. Maddala, C.R. Rao, and H.D. Vinod, eds., *Handbook of Statistics*, vol. 11, Elsevier, St. Louis, Mo., 455-488.

[278] Ohanian, L., Stockman, A. and Kilian, L. (1995), "The Effects of Real and Monetary Shocks in a Business Cycle Model with some Sticky Prices", *Journal of Money Banking and Credit*, 27, 1210-1240.

[279] Otrok, C. (2001), "On Measuring the welfare costs of Business Cycles" *Journal of Monetary Economics*, 47, 61-92.

[280] Otrok, C. and Whitemann, C. (1998), "Bayesian Leading Indicators: measuring and Predicting Economic Conditions in Iowa", *International Economic Review*, 39, 997-1114.

[281] Pagan, A. (1994), "Calibration and Econometric Research: An Overview", *Journal of Applied Econometrics*, 9, S1-S10.

[282] Pagan, A. and Harding, D. (2002), "Dissecting the Cycle: A Methodological Investigation", *Journal of Monetary Economics*, 49, 365-381.

[283] Pagan, A. and Shannon, J. (1985), "Sensitivity Analysis for Linearized Computable General Equilibrium Models", in J.Piggott and J. Whalley (eds.) *New Developments in Applied General Equilibrium Analysis*, Cambridge University Press, Cambridge, UK.

[284] Pagan, A. and Yoon, Y., (1993), "Understanding Some Failures of Instrumental Variable Estimators", University of Rochester, manuscript.

[285] Pappa, P. (2003), "New Keynesian or RBC transmission? The effects of fiscal shocks in labor markets", LSE and IGIER, manuscript.

[286] Pesaran, H. (1995) "Cross sectional aggregation of linear dynamic models", University of Cambridge, manuscript.

[287] Pesaran, H., Shin, Y. and Smith, R. (1999) "Pooled Mean Group Estimation of Dynamic Heterogeneous Panels", *Journal of the American Statistical Association*, 94, 621-634.

[288] Pesaran, H. and Smith, R. (1992), "The Interaction between Theory and Observation in Economics", University of Cambridge, manuscript.

[289] Pesaran, H. and Smith, R. (1995) " Estimating Long Run Relationships from Dynamic Heterogeneous Panels ", *Journal of Econometrics*, 68, 79-113.

[290] Pesaran, H., Smith, R. and Im, K. (1996) "Dynamic Linear Models for Heterogeneous Panels", in L. Matyas and P. Sevestre (eds.) *Econometrics of Panel Data*, Kluwer, New York, NY.

[291] Poirier, D. (1995) *Intermediate Statistics and Econometrics*, MIT Press, Cambridge, Ma.

[292] Press, A. (1997), *Numerical Recipes: The Art of Scientific Computing*, Cambridge University Press, Cambridge, UK.

[293] Priestley, I. (1981), *Spectral Analysis and Time Series*, Academic Press, Burlington, Ma.

[294] Proietti, T. and Harvey, A. (2000), "The Beveridge Smoother", *Economic Letters*, 67, 139-146.

[295] Quah, D. (1990), "Permanent and Transitory Movements in Labor Income: An Explanation for Excess Smoothness in Consumption ", *Journal of Political Economy*, 98, 449-475.

[296] Quah, D. (1996), "Regional Convergence Cluster across Europe", *European Economic Review*, 40, 951-958.

[297] Rao, C. R. (1973), *Linear Statistical Inference and Its Applications*, Wiley & Sons, Chichester, UK.

[298] Rao, C. R. (1975), "Simultaneous estimation of Parameters in Different Linear Models and applications to Biometric Problems", *Biometrics*, 31, 545-554.

[299] Ravn, M. and Uhlig, H. (2002), On adjusting the HP filter for the frequency of Observations, *Review of Economics and Statistics*, 84, 371-375.

[300] Ripley, B. (1987), *Stocastic Simulations*, Wiley & Sons, New York, NY.

[301] Robert, C. and Casella, G. (1999), *Monte Carlo Statistical Methods*, Springer and Verlag, New York, NY.

[302] Robertson, J. and Tallman, E. (1999), 'Vector Autoregressions: Forecasting and Reality", *Federal Reserve Bank of Atlanta, Economic Review*, First quarter, 4-18.

[303] Rose, A. (2004), "A Meta-analysis of the Effects of Common Currencies on International Trade", NBER working paper 10373.

[304] Roll, R. and Ross, S. (1980), " An empirical Investigation of the Arbitrage Pricing Theory, *Journal of Finance*, 35, 1073-1103.

[305] Rosenblatt, A. (1978), Dependence and Asymptotic Dependence for Random Processes, Rosenblatt, A. (ed.) *Studies in Probability Theory*, Washington, D.C.

[306] Ross, S. (1976), "The Arbitrage Theory of the Capital Asset Pricing", *Journal of Economic Theory*, 13, 341-360.

[307] Rotemberg, J. (1984), "Monetary equilibrium model with Transaction Costs", *Journal of Political Economy*, 92, 40-58.

[308] Rotemberg, J. and Woodford, M. (1997), "An Optimization Based Econometric Framework for the Evaluation of Monetary Policy", *NBER Macroeconomic Annual*, 12, 297-346.

[309] Rozanov, Y., (1967), *Stationary Random Processes*, Holden Day, San Francisco, Ca.

[310] Ruge-Murcia, F. (2002), Methods to Estimate Dynamic Stochastic General Equilibrium Models, UCSD working paper 2002-46.

[311] Runkle, D. (1987), "Vector Autoregression and Reality", *Journal of Business and Economic Statistics*, 5, 437-442.

[312] Sala, X., Doppelhofer, G., Miller, R. (2003), "Determinants of Long Term growth: A Bayesian Averaging of Classical Estimates (BACE) approach", Columbia University and Universitat Pompeu Fabra, manuscript.

[313] Sargent, T. (1979), "A Note on Maximum Likelihood estimation of Rational Expectations Model of the Term Structure", *Journal of Monetary Economics*, 5, 133-143.

[314] Sargent, T. (1986), *Dynamic Macroeconomic Theory*, Harvard University Press, Cambridge, Ma.

[315] Schorfheide, F. (2000), "Loss function based evaluation of DSGE models", *Journal of Applied Econometrics*, 15, 645-670.

[316] Serfling, R, (1980), *Approximation theorems of Mathematical Statistics*, Wiley & Sons, Chichester, UK.

[317] Shapiro, M. and Watson, M. (1988), "Sources of Business Cycle Fluctuations", *NBER Macroeconomic Annual*, 3, 111-148.

[318] Showen, J. and Whalley, J. (1984), "Applied General Equilibrium Models of Taxation and International Trade: An Introduction and Survey", *Journal of Economic Literature*, 22, 1007-1051.

[319] Showen, J. and Whalley, J. (1992), *Applying General Equilibrium*, Cambridge University Press, Cambridge, Ma.

[320] Simkins, S.P. (1994), "Do Real Business Cycle Models Really Exhibit Business Cycle Behavior?", *Journal of Monetary Economics*, 33, 381-404.

[321] Sims, C. (1980), "Macroeconomics and Reality", *Econometrica*, 48, 1-48.

[322] Sims, C. (1988), " Bayesian Skepticism on unit root econometrics", *Journal of Economic Dynamics and Control*, 12, 463-474.

[323] Sims, C. (1996), "Macroeconomics and Methodology", *Journal of Economic Perspectives*, 10, 105-120.

[324] Sims, C. (2001), Drift and Breaks in Monetary Policy, Princeton University, manuscript.

[325] Sims, C. and Sargent, T. (1977), "Business Cycle Modelling without Pretending to have too much a priori economic theory" in *New Methods in Business Cycle Research*, Federal Reserve Bank of Minnneapolis, Minneapolis, MN, 45-109.

[326] Sims, C., Stock, J. and Watson, M. (1990), "Inference in Linear Time Series Models with some unit roots", *Econometrica*, 58, 113-144.

[327] Sims, C. and Zha T. (1998), "Bayesian Methods for Dynamic Multivariate Models", *International Economic Review*, 39, 949-968.

[328] Sims, C. and Zha, T. (1999), "Error Bands for Impulse Responses", *Econometrica*, 67, 1113-1155.

[329] Sims, C. and Zha, T. (2004), "Macroeconomic Switching", Federal Reserve of Atlanta, working paper 2004-12.

[330] Smets, F. and R. Wouters (2003), " An Estimated Stochastic DGE model of the Euro Area", *Journal of the European Economic Association*, 1, 1123-1175.

[331] Smith, A.F.M. (1973), "A General Bayesian Linear Model", *Journal of the Royal Statistical Society*, ser B, 35, 67-75.

[332] Smith, A.F.M. and Roberts, G.O, (1993), "Bayesian Computation via the Gibbs sampler and related Markov Chain Monte Carlo methods" *Journal of the Royal Statistical Society, B,* 55, 3-24.

[333] Smith, A. (1993), "Estimating Nonlinear Time Series Models using Simulated Vector Autoregressions", *Journal of Applied Econometrics*, 8, 63-84.

[334] Smith, R. and Fuertes, A. (2003), "Panel Time Series", Birbeck College, manuscript.

[335] Söderlind, P. (1994), "Cyclical Properties of a Real Business Cycle Model", *Journal of Applied Econometrics*, 9, S113-S122.

[336] Sorensen, B., Wu, L. and Yosha, O. (2001), "Output fluctuations and fiscal policy: US state and local governments, 1978-1994, *European Economic Review*, 45, 1271-1310.

[337] Stock J. and M. Watson (1987), "Testing for Common Trends", *Journal of the American Statistical Association*, 83, 1996-1107.

[338] Stock, J and Watson, M. (1989), "New Index of Coincident and Leading Indicators, *NBER Macroeconomics Annual*, 4, 351-396.

[339] Stock J. and Watson, M. (1991), "A Probability Model of the Coincident Economic Indicators" in Lahiri, K. and Moore, G. (eds.) *Leading Economic Indicators: New Approaches and Forecasting Records*, Cambridge University Press, Cambridge, UK.

[340] Stock, J. and Watson, M (2002), *Econometrics*, Addison Wesley, Boston, Ma.

[341] Stout, W., (1974), *a.s. convergence*, Academic Press, Burlington, Ma.

[342] Swamy, P. (1970), " Efficient Inference In a Random Coefficients Regression Model", *Econometrica*, 38, 311-323.

[343] Tamayo, A. (2001), "Stock Return predictability, Conditional Asset Pricing Models and Portfolio Selection", University of Rochester, manuscript.

[344] Tanner, M. and Wong, W.H. (1987), "The Calculation of Posterior Distributions by Data Augmentation" , *Journal of the American Statistical Association*, 82, 528-550 (with discussion).

[345] Tauchen, G., (1986), "Statistical Properties of GMM estimators of Structural Parameters obtained from Financial Market Data", *Journal of Business and Economic Statistics*, 4, 397-425.

[346] Tierney, L (1994), "Markov Chains for Exploring Posterior Distributions (with discussion), *Annals of Statistics*, 22, 1701-1762.

[347] Uhlig, H. (1997), "Bayesian Vector Autoregression with Stochastic Volatility", *Econometrica*, 65, 59-73.

[348] Uhlig, H. (1999), "What are the Effects of Monetary Policy? Results from an agnostic Identification procedure", Tilburg University, working paper.

[349] Uhlig, H. (1999), " A methods for Analyzing Nonlinear Dynamic Stochastic Models Easily" in Marimon, R. and Scott A. (eds.) *Computational Methods for the Study of Dynamic Economies*, Oxford University Press, Oxford, UK., 114-142.

[350] Zellner, A. (1971), *Introduction to Bayesian Inference in Econometrics*, Wiley & Sons, Chichister, UK.

[351] Zellner, A., Hong, (1989), "Forecasting International Growth rates usign Bayesian Shrinkage and other procedures", *Journal of Econometrics*, 40, 183-202.

[352] Zha, T. (1999), "Block Recursion and Structural Vector Autoregressions", *Journal of Econometrics*, 90, 291-316.

[353] Wabha, G. (1980), "Improper prior, Spline smoothing and the Problem of guarding against models errors in regression", Journal of the Royal Statistical Association, ser B, 40, 364-372.

[354] Waggoner, D. and Zha, T. (1999), "Conditional Forecasts in Dynamic Multivariate Models", *Review of Economics and Statistics*, 81, 1-14.

[355] Waggoner, D. and Zha, T. (2003), "A Gibbs Simulator for Restricted VAR models", *Journal of Economic Dynamics and Control*, 26, 349-366.

[356] Watson, M. (1986), "Univariate Detrending Methods with Stochastic Trends", *Journal of Monetary Economics*, 18, 49-75.

[357] Watson, M. (1989), "Recursive Solution Methods for Dynamic Linear Rational Expectations Models", *Journal of Econometrics*, 41, 65-89.

[358] Watson, M. (1993) "Measures of Fit for Calibrated Models", *Journal of Political Economy*, 101, 1011-1041.

[359] Watson, M., (1994), "Business Cycle Duration and Postwar stabilization of the Us economy", *American Economic Review*, 84, 24-46.

[360] Watson, M. (1995), "VAR and Cointegration" in Engle, R. (ed.), *Handbook of Econometrics, IV*, Elsevier, St. Louis, Mo.

[361] Wei, Y. (1998), "Can a RBC model pass the Watson test? *Journal of Monetary Economics*, 42, 185-203.

[362] West, K. (1995), "Another Heteroskedastic and Autocorrelation Consistent Covariance Matrix Estimator", NBER, technical working paper 183.

[363] West, K. and Wilcox, D. (1996), "A Comparison of Alternative Instrumental Variables Estimators of a Dynamic Linear Model", *Journal of Business and Economic Statistics*, 14, 282-294.

[364] White, H. (1982), "Maximum Likelihood Estimation of Misspecified Models", *Econometrica*, 50, 1-25.

[365] White, H., (1984), *Asymptotic Theory for Econometricians*, Academic Press, Burlington, Ma.

[366] White, H. and Domowitz, I. (1984), "Nonlinear Regression with Dependent Observations", *Econometrica*, 52, 143-161.

[367] Whittle, P. (1990), *Prediction and Regulation*, University of Minnesota Press, Minneapolis, Mn.

[368] Wilcox, D. (1992) " The Construction of US Consumption Data: Some Facts and their Implication for empirical work, *American Economic Review*, 82, 922-841.

[369] Wolman, A. (2001), "A Primer on Optimal Monetary Policy with Staggered Price Setting", *Federal Reserve Bank of Richmond, Economic Review*, Fall, 87/4, 27-52.

[370] Woodford, M. (1996), "Control of the Public Debt: A Requirement for Price Stability?", NBER working paper 5684.

[371] Woodford, M. (2003), *Interest and Prices: Foundation of a theory of Monetary Policy*, Princeton University Press, Princeton, NJ.

Index