ESSAYS IN ECONOMETRICS AND FORECASTING

Nicholas Fawcett

NEW COLLEGE

UNIVERSITY OF OXFORD

A THESIS

SUBMITTED FOR THE DEGREE

OF DOCTOR OF PHILOSOPHY AT THE

UNIVERSITY OF OXFORD

HILARY TERM 2008

TO NICKY

Contents

List of Figures

List of Tables

Preface

Unpredictable breaks, as I have found when writing this thesis, need not be confined to the rate of inflation, or a weather forecast. The route that this work has taken has been subjected to many unexpected events and developments, and the fact that the final product bears little resemblance to my initial ideas at the start of the M.Phil. serves as a reminder (if one were necessary) that breaks happen all the time.

Fortunately I have had a robust strategy of sorts, in the form of excellent teaching and supervision over the past few years. My first thanks are therefore due to my supervisors, David Hendry, and more recently Steve Bond. David's teaching and careful guidance over the course of the M.Phil. and D.Phil. has been invaluable, and helped improve me immeasurably as an econometrician. Steve provided expert advice, and I am grateful for his support, especially when there was a tight timetable to work within. Both have helped shape this thesis for the better.

The task of doing research and talking about it each week was made all the more enjoyable by the other members of David's research group, and so I am pleased to thank Jennie Castle, Julia Giese, Patrick Lo, James Reade and Christin Kyrme Tuxen for their suggestions and comments. James in particular gave freely of his time, and of his vast library of Ox code, and I am very grateful to him.

Over the past few years, I have accumulated a large debt of gratitude to seminar and conference participants, and I have pleasure here in acknowledging the suggestions of Clive Bowsher, Mike Clements, David Greenstreet, Siem Jan Koopman, Clare Leaver, Dilek Onkal, Kevin Sheppard, Francis Teal and Ken Wallis. Further thanks are also due to Joerg Breitung, Jurgen Doornik, Markus Eberhardt, Jan Groen, James Mitchell, Bent Nielsen, Alisdair Scott, Adrian Pagan, M. Hashem Pesaran, Simon Price, and Simon Wren-Lewis. Andrew Glyn and David Walton provided extremely helpful guidance during the early stages of my work on trade and forecast evaluation respectively; they are both greatly missed.

Aside from the confines of academic work, I have found support from many excellent friends, including Mark Abrahamson, Sudhir Anand, Simon Baptist, Sarah Cochrane, John William Devine, Andrew Elliott, Justin Fansler, Dave Foster, Michael Griebe, Rudy Kleysteuber, Dirk-Jan Omtzigt, Joe Perkins, Jari Stehn, and Bob Rijkers. Though lost on Dirk-Jan, LaTeX

has been invaluable in writing this thesis, and I have pleasure in thanking Thomas Fink for his excellent style files, which he will no doubt recognise throughout the chapters.

For inspiring my interest in the subjects I have explored, I am delighted to thank Nigel Allington, Michelle Baddeley, Michael Kitson, and Iain Macpherson; and for showing me what economics could do, I thank Roger Loxley, Geoff Riley and Peter Shelley.

Funding from the Economic and Social Research Council under award PTA-030-2003-00904 was indispensable. In the later stages of the D.Phil., I was fortunate, and not to mention grateful, to receive further support through a Doctoral Studentship from the Oxford Department of Economics and St Catherine's College. The Webb-Medley fund in the Economics Department, and New College, both provided financial assistance to attend conferences. As a student, I have had the pleasure of studying at two great colleges: first Brasenose, during my M.Phil., and New College during the D.Phil. As a tutor, St Catherine's College provided a warm and welcoming environment to start an academic career in.

Although Gavin Cameron could not see the final product, if he had read through it he would have found ample evidence of his influence on my approach to doing research, and writing about it. He was a great support, a gifted teacher, and an inspiration.

Closer to home, I am extremely grateful to my parents, whose continued support and generosity allowed me to pursue this thesis, and Julia and Ben for entertaining times whilst it was being written. Above all, I thank Nicky for her patience, kindness and encouragement; this thesis could not have been written without her.

Nicholas Fawcett
New College
Oxford
May 2008

Abstract and Word Count

Whether we would like to model imports and exports, or forecast inflation, structural variation in an economy frequently causes problems. This thesis examines such variation in two dimensions: first, in a cross-section of individuals, and secondly, over time. A panel of manufacturing industries in several developed countries reveals that there is substantial variation across sectors, in the response of trade to changes in prices and incomes. Ignoring this heterogeneity can render conventional results biased and inconsistent, so a number of robust methods are used to obtain reliable estimates of long-run and short-run trade relationships. The findings point to common behaviour across sectors, which could be due to similarities in technology.

The impact of structural breaks over time is examined in the second part of the thesis. Unpredictable shifts in deterministic terms such as the mean of a process are shown to generate significant forecast failure, and even the methods used to evaluated forecast accuracy are affected. Using the Kullback-Leibler discrepancy to measure the size of forecast errors, various robust mechanisms are discussed, that do not fail systematically after a break. Although they can provide a degree of insurance if a shift does occur, this comes at a cost if there is no change, and in the presence of measurement error they can exacerbate the uncertainty surrounding a forecast. An empirical illustration with a model of UK money demand provides some support for the automatic correction mechanisms, although there does seem to be a role for direct modeling of a break process.

JEL Keywords: C32, C52, C53, F14

The body of this thesis comprises 197 numbered pages, with a typical page of text containing around 376 words, making the document approximately 74,000 words long.

Several applications were used to produce the thesis, figures and calculations: the text was set using LaTeX, and the figures with PicTeX; *Ox* and *GAUSS* were used for computations and some simulations; *PcNaive* produced the

Ox code for the remaining simulations; *PcGive* was used for estimation and forecasting; and *Mathematica* was used for discrepancy calculations (including numerical integration) and two figures. The author would like to thank Joerg Breitung, M. Hashem Pesaran and James Reade for kindly providing *Ox* and *GAUSS* code for estimation. Full citations are provided in the text.

The LaTeX formatting was heavily influenced by Thomas Fink and his style package, `farrfink.tex`, to whom the author is deeply grateful.

Nomenclature

General Econometric and Statistical terms

$\mathsf{P}(\mathcal{A})$	the probability of event $\mathcal{A}$ occurring
$\mathsf{D}_{\mathsf{X}}(x)$	probability density function for a random variable X, evaluated at x
$\mathsf{F}_{\mathsf{X}}(x)$	cumulative density function for a random variable X, evaluated at x; thus $\mathsf{F}_{\mathsf{X}}(x) = \mathsf{P}(X \leq x)$
$\mathsf{E}(X)$	expectation of a (random) variable X
$\mathsf{E}(X\|Z)$	conditional expectation of X, conditional on another (possibly vector) variable Z
$\mathsf{V}(X)$	variance of a random variable X
$\mathsf{M}(X)$	mean-squared error of a random variable, i.e. $\mathsf{M}(\cdot) = \mathsf{V}(\cdot) + (\mathsf{E}(\cdot))^2$
$\mathsf{NID}(\mu, \sigma^2)$	normally and independently distributed with mean μ and variance σ^2; if the series is not independent, then the distribution is denoted $\mathsf{N}(\mu, \sigma^2)$.
$\mathsf{IN}_k(\boldsymbol{\mu}, \boldsymbol{\Omega})$	k-dimensional (independent) multivariate normal distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Omega}$
$\mathsf{UID}(a, b)$	uniformly and independently distributed on a support from a to b; lack of independence is indicated by U
IID	independently and identically distributed
$\phi(z)$	standard normal density function evaluated at z
$\Phi(z)$	standard normal distribution function evaluated at z
$\chi^2(v)$	Chi-squared distribution with v degrees of freedom
$\stackrel{\mathsf{D}}{=}$	equality in distribution
$\mathbb{R}^k$	k-dimensional set of real numbers
$\mathbf{z}$, $\mathbf{x}$ etc	vectors of variables (denoted by bold face)
$\mathbf{I}_n$	$n \times n$ identity matrix
$\mathsf{I}(d)$	integrated of order d: see Banerjee *et al.* (1993, p. 84)
DGP	data generating process
OLS	Ordinary least squares estimator
2SLS	Two-stage least squares estimator
GMM	Generalised Method of Moments estimator

Panel Estimation of Trade Elasticities

CCE	Common Correlated Effects estimator – see Pesaran (2006)
MGOLS	Mean Group OLS estimator

Forecasting terms

VEqCM	Vector Equilibrium Correction Mechanism
DDD	Double-Differenced Device
DVEqCM	Differenced Vector Equilibrium Correction Mechanism
$\mathsf{f}_{\mathsf{Y}_t}(y_t)$	true density function of a random variable y at time t
$\mathsf{g}_{\mathsf{Y}_{t+h},t}(y_{t+h})$	h-step ahead forecast denstity of y at time $t+h$; produced at time t
$\nabla\boldsymbol{\mu}^*$	change in the $(r \times 1)$ equilibrium mean: $\nabla\boldsymbol{\mu}^* = \boldsymbol{\mu}^* - \boldsymbol{\mu}$, where the $*$ denotes the post-break value
$\widehat{\mathbf{x}}$	VEqCM forecast of $\mathbf{x}$
$\widetilde{\mathbf{x}}$	DDD forecast of $\mathbf{x}$
$\widetilde{\widetilde{\mathbf{x}}}$	DVEqCM forecast of $\mathbf{x}$
I_{KL}	Kullback-Leibler (KL) relative entropy (equivalently called the KL discrepancy, information criterion/statistic, distance)
I_{SE}	Integrated squared error, measure of discrepancy

Forecast-Error Correction

σ^2_ϵ	Measurement error variance
σ^2_η	DGP innovation variance (in the dynamic model)

Money Demand models

$\mathbb{M}_{\mathsf{R}_{LA}}$	Estimated system using the local authority interest rate R_{LA}
$\mathbb{M}_{\mathsf{R}_D}$	Estimated system using the interest rate differential R_D
$\mathbb{M}_{\mathsf{R}_S}$	Estimated system using the learning-adjusted interest rate R_S

Density Evaluation

$\{z_t\}_{t=1}^n$	sequence of probability integral transforms from time $t=1$ to n
$\{z_t^*\}_{t=1}^n$	sequence of probability integral transforms mapped onto the quantiles of a normal distribution, from $t=1$ to n.
$L(a_t, y_t)$	loss function of an agent, for a given action choice a_t and realisation y_t
DH	Doornik-Hansen test for normality (see Doornik and Hansen (1994))

ESSAYS IN ECONOMETRICS AND FORECASTING

Chapter 1

Introduction

> Earlier on today apparently a lady rang the BBC and said she heard that there was a hurricane on the way. Well don't worry if you're watching, there isn't.
>
> MICHAEL FISH, BBC WEATHER FORECASTER
> 15 October 1987

> Hurricane winds batter southern England
>
> BBC NEWS HEADLINE
> 16 October 1987

Whether we would like to model imports and exports, or forecast inflation, structural variation in an economy frequently causes problems. In some cases, it is manifest in changes over time, but in others the differences emerge across groups within an economy, such as individual industries. This thesis explores how we might build models of trade flows or inflation, that take these phenomena into account.

Why is it important to do so? From an econometric point of view, the motivation is clear. Forecast models can suffer from systematic failure if they are not robust to structural breaks, and even the methods used to evaluate the accuracy of a forecast can break down. Similarly, a model that ignores differences across individual elements by focussing only on an aggregate picture, may yield unreliable results.

A policy perspective suggests that it is essential to have a forecasting and modeling strategy that can deliver reliable predictions, if the forecasts are used as an intermediate input in a wider decision-making process. Furthermore, policies that are aimed at specific parts of the economy may have

differing effects if each part does not behave in the same way, so well-directed policy requires well-specified estimates.

Given the myriad forms of structural differences, it is clear that the problems faced when trying to respond to these issues are substantial, so the approach taken here is to divide the overall question into two parts.

The first concentrates on heterogeneity in a cross-section dimension, by asking whether there are differences across countries and industries in the way trade flows respond to changes in prices and income. Thus Chapter 2 uses data on a panel of manufacturing industries in a set of developed countries (members of the Organisation of Economic Co-operation and Development, or OECD) to estimate price and income elasticities for imports and exports. In doing so, the chapter explores whether there are similarities in the way a particular industry behaves in several countries, or in the way all industries in a particular country behave, against the idea that *all* industries in *all* countries respond in the same way. As it turns out, the evidence supports the hypothesis that there are substantial differences in the long-run relationships across industries, and these are estimated in the chapter.

In the second part, a robust forecasting strategy is developed, which looks at the effect of structural breaks over time on two parts of a forecasting process: first, the stage of *production*, when the objective is to build a forecast model, and secondly, the stage of *evaluation*, where forecast accuracy is assessed retrospectively.

As Chapter 3 shows, unanticipated breaks can lead to serious forecast failure when they emerge via changes in deterministic elements such as the long-run mean of the random variable being forecast.[1] This idea draws on an established literature on forecasting, that points to such shifts as being the most important source of forecast failure (see, for example, Clements and Hendry (1998) or Pesaran, Pettenuzzo and Timmermann (2006)). For the class of equilibrium-correction models examined in this thesis, the intuition

[1]Forecast failure is typically defined as a significant worsening of forecast performance relative to anticipated outcomes.

behind this is straightforward: at any point in time an equilibrium-correction process tends towards its long-run mean value, and as long as a forecast model correctly estimates what this is, then on average the predictions it produces will be unbiased. If, however, there is a mean shift and the forecasts do not adapt to this, then there will be *systematic* forecast failure.

Precisely how bad this failure might be, is explored in Chapters 4, 5 and 6, which consider how models can be made robust to the effect of structural breaks.

Chapter 4 provides an overview of how this can be done, and highlights that robust models do not suffer from permanent forecast failure if a break occurs. Further, a robust forecasting *strategy* acknowledges that insuring a model against breaks, so that it is robust, can incur a cost in terms of increased uncertainty. Thus a strategic choice may exist, in which the cost of using a robust model must be balanced against the benefits of doing so when shifts do occur. In trying to quantify this trade-off, the chapter introduces discrepancy measures that can capture complete forecast densities, rather than the mean and variance alone (as the popular mean-squared forecast error does), and discusses why this might be relevant to some applications of forecast models in a decision-making environment.

Lest Chapter 4 give the impression that robust forecasting devices are a panacea for forecast failure due to breaks, Chapter 5 points out how some devices can exacerbate problems when there are measurement errors in the observations of a random process. Using an intercept-correction model, which includes lagged forecast errors when making predictions, the chapter demonstrates that the robust solution to structural breaks is totally at odds with the best response when there are errors in measurement. The intuition behind this is that a permanent structural break in the mean of a process should be *incorporated into* forecasts of its future value, whilst one-off measurement error should be *offset*, as it provides no genuine information about the process at the moment a forecast is made. In response, the chapter considers how lagged forecast errors could still be used to improve future forecasts, even if breaks and measurement error is likely, by

allowing the weight applied to each error to vary. Allowing for variable – and indeed negative – weights can improve forecast performance in theory, and the chapter discusses why this might be the case, for different degrees of break size and measurement uncertainty.

Drawing on the foregoing discussion of robust forecasting models, Chapter 6 applies some of the approaches to a real-world example, by studying the role of learning and adjustment in a model of UK money demand.[2] This revolves around the Banking Act of 1984, and although the episode has been studied extensively in the literature, the perspective in this chapter is new. The role of wider influences, including legislative changes, in generating structural breaks, is distinguished from economic factors that keep an economic system going in its post-break state, and the application to money demand forecasts shows how the learning or adjustment process in the aftermath of the Banking Act can be decomposed. Although robust mechanisms emerge the victor in comparisons of forecast accuracy in the early stages after the legislative change, a learning-adjusted model based on economic theory performs better as the economy completes its adjustment to a new equilibrium mean.

Besides using robust *models*, the methods of *evaluating* forecast accuracy should also remain unaffected if there are breaks. To see whether this is the case, Chapter 7 introduces a popular method of density evaluation and simulates its performance in a range of different break situations. The motivation for evaluating densities, rather than point expectations, comes from the fact that forecasts are frequently used as an intermediate input when making decisions. Probability densities play an important role in establishing the expected payoff from different decisions, when the variable being forecast can take many different values. To see this more concretely, consider the objective function of a monetary authority such as the Bank of England, which makes interest rate decisions on the basis of an inflation

[2]This chapter contains material that has been included in a joint paper with David F. Hendry and Jennifer L. Castle, which has been invited for submission to the *Journal of Econometrics*.

forecast, to achieve an objective of 2% inflation at a two-year horizon. Each inflation forecast takes the form of a complete probability density, which captures not only the most probable inflation outturn, but also the uncertainty around it, and this information is important when the Bank decides not only whether to change interest rates, but also by how much to change them The chapter finds that in simulation exercises, statistically small (but economically significant) breaks may not be detected, whilst large but *temporary* breaks are not distinguished from *permanent* shifts. An application of the evaluation method to the Bank of England's inflation forecasts finds evidence of accurate forecasts at close horizons, but structrual breaks in forecasts over longer intervals.

Part 1

HETEROGENEITY AND NON-STATIONARITY IN TRADE ELASTICITIES

Chapter 2

Trade, Income and the Exchange Rate in the OECD

> More and more of our imports come from overseas
>
> US PRESIDENT GEORGE W. BUSH

How responsive are imports and exports to changes in relative prices and income? On a country-wide level, there is a large literature that answers the question. However, given differences among the kind of sectors that make up an economy, it seems likely that such a broad picture will disguise substantial heterogeneity in the trading behaviour of individual industries. The aim of this chapter, therefore, is to explore the price and income elasticities of imports and exports in a disaggregated panel of industries across a number of countries in the Organisation for Economic Co-operation and Development (OECD).

Using a dataset covering thirteen manufacturing industries in eleven countries, similarities in the behaviour of sectors across countries are exploited to obtain estimates of long-run trade relationships. The broad picture that emerges from this analysis is that there is significant variation in elasticities across industries.

The motivation for studying this topic at an industry level comes from the models of innovation, trade and growth developed by Cameron, Proudman and Redding (2005) and Redding (2002), amongst others. They emphasise that trade performance is related to the level of technological progress in an economy, and that countries can develop competitive trading advantages over competing economies by innovating enough to develop a technological lead over rivals. This raises two issues. First, there is no reason to assume

that all the industries in an economy are at the same level of technological leadership, relative to others; thus the UK could be a leader in pharmaceutical products, for example, but not in furniture making. As a result, policies aimed at raising innovation expenditure, and ultimately improving long-run growth, might be more effective in some industries than others.

Secondly, temporary shocks in trade performance – due to the exchange rate, for instance – might have a long-term impact on some sectors, if expenditure on innovation and future technological improvements are affected.

A disaggregated analysis of trade performance across industries is relevant to these issues because it can quantify differences in industry behaviour, in the face of changes in prices, income and other factors (such as innovation). It also offers a framework to study the long-run impact of short-run shocks.[1] Since there is relatively little existing work on this level of disaggregation, using the OECD database discussed here, the analysis in this chapter represents the beginning of a wider research agenda. The scope of empirical work is restricted to examining only price and income elasticities, for a subset of manufacturing industries, over a short period of time. With this assessment accomplished, work can then proceed on the wider issues discussed above.

2.1 Econometric context

The approach taken in this chapter represents a contribution to the literature in terms of the scope of the empirical work, so it is useful first to provide a *tour d'horizon* of the literature on trade elasticities before moving to the details of the estimation exercise in hand.

In considering how the existing literature relates to this chapter, it is helpful to distinguish between the state of the literature as Goldstein and Khan (1985) saw it in their influential review, and subsequent developments.[2] The rationale for this lies in the fact that the methodological ap-

[1] In the trade literature, these phenomena are studied in models of hysteresis: see Baldwin (1988) and Baldwin and Lyons (1994).

[2] In doing so, the Goldstein and Khan review is taken to represent the 'state of the art' in this literature at the time of its writing.

proach in the earlier period suffered from problems caused by the incorrect treatment of non-stationarity, and this has only been addressed, in the later literature, with econometric techniques that were themselves developed in the 1980s.

2.1.1 The early trade literature

Looking at the earlier literature, there are a number of points to highlight that are still relevant now, despite the methodological problems outlined above. Three broad considerations are important: first, the extent to which imported or exported goods are perfectly substitutable for domestically-produced alternatives; secondly, based on this, the choice of appropriate variables to include in empirical models; and finally, the degree of heterogeneity allowed for across the sectors in the economy.

In terms of substitutability, the scope of this study provides a clear indication of whether trade flows are perfect substitutes for domestic goods. The objective in this chapter is to examine the effects of income and prices on trade in a range of *manufacturing* sectors. Even though the level of disaggregation within industries is quite high, the goods themselves are sufficiently differentiated that it seems unreasonable to believe that they could be perfect substitutes for domestic goods. Further, as Goldstein and Khan (1985, p. 1044-45) point out, under perfect substitutability, one might expect first, either a domestic *or* foreign good swallowing an entire market (if each is produced with constant or decreasing costs), and secondly, a country to be an importer of a good or an exporter, but not both. These predictions are clearly violated in practice, so lending support to a model of imperfect substitutes.

The resulting set of equations that comprise a general imperfect substitutes model is presented by Goldstein and Khan (1985, p. 1045), and so the focus here is on the key relationships that underpin the econometric

modelling strategy below. They can be expressed generically as:

$$\mathsf{X}_i = f(\mathsf{RPX}_i, \mathsf{YF}_i) \tag{2.1a}$$

$$\mathsf{M}_i = g(\mathsf{RPM}_i, \mathsf{YD}_i), \tag{2.1b}$$

for some functions $f(\cdot)$ and $g(\cdot)$, where for country i: X_i and M_i are real exports and imports respectively; RPX_i and RPM_i are the relative prices of exports and imports against an index of the prices of substitute goods; and YD_i and YF_i represent domestic and foreign income respectively.

Against such a general background, further detail is needed on the variables in (2.1), which leads to the second issue raised above, on the choice of measures to use. This revolves around two questions: how to construct suitable price indices for use in deflating the nominal trade flows, and the relative price measures; and the nature of the income term that should be included.

The choice of import and export price indices is relevant to both the dependent variables X_i and M_i, and the relative price terms themselves. With regard to the former, as Goldstein and Khan (1985, p. 1054) observe, raw trade data are frequently provided in current prices (i.e. value rather than volume terms), and so an appropriate deflator is needed to transform the value series into a quantity measure. In principle, an ideal deflator would capture the prices of exactly those goods exported or imported, but in practice such information is rarely recorded for more than an extremely limited range of countries and time periods.[3] Consequently, current price values must be deflated using wider measures of prices such as wholesale price indices or unit value measures. This approach does have its limitations; for example, when taken across an entire country, the notion of a representative 'unit' of exports or imports does not sit well when the products that comprise either measure may include items as diverse as tractors, computers and financial services. Further, where price indices include tradable and

[3]Goldstein and Khan point to the NBER export price series which do capture this information, but which only cover four countries, and only for short spans across the 1950s and early 1960s.

nontradable goods, divergence in the pricing behaviour of the two groups could bias a deflator, which should only be adjusted for price changes in the former group. The overall effect of these elements could be to introduce measurement error in both dependent and independent variables, if the import or export price series used to deflate value series were also used in the construction of relative price measures. When looking at an aggregate (say, economy-wide) relationship, however, alternative measures may not exist; thus as Goldstein and Khan (p. 1056) reflect, the most pragmatic approach may be to "accept the danger of biased estimates, and exercise due caution on the ranges of the true elasticities." A more active response to the problem is pursued later in this chapter, since disaggregated data are used to help attenuate the impact of aggregation on price indices: this is discussed in Section 2.2.

Besides the choice of import or export price measures, the other component of the relative price, namely the price of substitute goods, needs to be considered. Theory would suggest including direct substitutes to traded goods, which highlights a distinction, discussed by Goldstein and Khan (1985, p. 1062), between domestically produced tradables and non-tradable goods. They point towards a two-step decision process, in which consumers first decide how much of their expenditure to allocate on tradables and non-tradables, based on the relative (overall) price of the two groups, and then *given* this, they then decide between domestic tradables and imported alternatives. This suggests that the relative price that should enter either the import or export equations is the ratio of import or export prices and the price of domestic tradables. Unfortunately, price data on the latter is not available, and a common response to this is to use an overall measure of prices such as the GDP deflator as an alternative. However, since a whole-economy measure is likely to comprise a significant proportion of *non-tradables* then the end result is that data constraints allow the price of non-tradables back into the *empirical* equation, even if they do not have a role in *theory*.

The choice of income measure, which represents the other element in

demand specification (2.1), differs slightly when considering imports or exports. Underlying either equation is a desire to capture the direct effect of greater income on demand. This suggests that an appropriate measure for use in the import model captures domestic income, and in the export model, captures income in the rest of the world. Within this framework, there is some discussion of whether income terms should be decomposed into trend and cyclical components (see Goldstein and Khan, p. 1057). The rationale for this question is that demand for imports or exports might be affected by the position of domestic or world income in their respective economic cycles, due to excess demand and constraints on domestic productive capacity. Thus in a situation of excess demand, where domestic production might be close to its potential level, demand for imports could increase, in the face of rising costs of production for domestic alternatives. This affects exports in a slightly different way, since there could be a theoretical role for domestic trend income (as a measure of productive capacity) entering into the export equation *alongside* overall world income. However, the bearing of such questions will depend on the frequency and span of data, and on whether domestic income would provide an adequate measure of productive capacity.[4]

Having considered the degree of substitutability of traded goods, and the choice of variables to include, the third point that arises from the early trade literature focuses on the level of aggregation allowed for in empirical studies. Although this is only briefly touched upon by Goldstein and Khan, it has an important bearing on later work in this chapter, and as such it is useful to highlight it now. Starting from a whole-economy viewpoint as the highest level of aggregation, a first step towards disaggregation could be according to broad commodity group (e.g. agricultural trade, raw materials, manufactures, etc), or by trade partner. This approach is taken by Houthakker and Magee (1969), who analyse trade elasticities by looking first

[4]Balassa (1979), for example, suggests that trend real income provides a poor measure of supply, since it assumes the same elasticity of substitution between exports and tradables, and exports and non-tradables.

at country-specific elasticities, then bilateral trade flows between the US and major partners, and finally US elasticities for five commodity groups.

It is worth highlighting two aspects of a disaggregated picture, as they are relevant to the later sections in this chapter. First is the ability to allow for heterogeneity in the trade flows across sectors in the economy, which recognises the fact that two sectors – say textiles and IT – might not share the same trading partners. This could be important, as the appropriate relative price and income measures for each sector need not be the same in this case: if the bilateral exchange rate with the domestic textile sector's major partner appreciated, whilst that of the IT sector depreciated, then relative prices could diverge. Similarly, income shocks affecting one sector's trading partners more than another's could generate the same kind of asymmetry with regard to income.

Secondly, it is possible that the import and export functions themselves – i.e. $f(\cdot)$ and $g(\cdot)$ from (2.1) – might vary by sector and country. In practice this would probably be manifest in heterogeneity across price and income elasticities, rather than a completely different functional form, but the differences could still be significant.

2.1.2 Recent developments

The foregoing analysis has identified several elements of an empirical modelling strategy that are as relevant to any current analysis as they were in the early trade literature, in which trade flows can be assessed in terms of relative prices and income.

However, a criticism of this early literature is that its treatment of non-stationarity in empirical models was poor. Whilst the stationarity of all variables was implicitly assumed, when they were included in *level* form in models, in reality this assumption later turned out to be invalid. As discussed below, tests for the presence of unit roots suggest that the variables in (2.1) are, in most cases, I(1), yielding potentially spurious results from a simple linear regression model including only the levels of each.[5]

[5]If the variables in a regression in levels are cointegrated, then superconsistency proper-

Houthakker and Magee (1969) provide a useful example of how early indications of this problem were addressed. Using log-linear specifications for the functions $f(\cdot)$ and $g(\cdot)$ in (2.1), they estimate equations for US imports and exports as a function of relative prices and income via OLS. As they note, the adjusted-R^2 statistic for all country trade equations (with the exception of South Africa and Australia) are high, ranging from 0.933 to 0.997, whilst the Durbin-Watson statistics are typically very low. Attributing this to an inadequate treatment of dynamics in the model, their response is to use an iterative Cochrane-Orcutt procedure to incorporate an estimate of the residual autocorrelation in the equations.

Goldstein and Khan pursue this approach to incorporating dynamics in models specified in terms of levels, discussing the alternative merits of Koyck lag specifications and Almon polynomial lags. However, none of the discussion in either study makes reference to the non-stationarity of the data used, or possible approaches for dealing with it. Indeed, reflecting on the problems that were systematic in the literature, Goldstein and Khan observed that "most of the major methodological pitfalls in the specification and estimation of trade models that currently preoccupy researchers had already been identified by the early 1950s" (p. 1097).[6]

The major methodological development that offered a solution to this problem came, of course, with the theory of cointegration in papers by Granger (1981) and Engle and Granger (1987) amongst others. In suggesting that there could be cases in which linear combinations of $\mathsf{I}(1)$ variables were $\mathsf{I}(0)$, this theory provided an appealing approach to the estimation of more reliable trade elasticities.[7] To see how this relates to the trade equations in (2.1), consider an example using imports, in which there is one cointegrating vector, and the long-run relationship is written in log-linear

ties of a coefficient estimator might provide meaningful point estimates of the coefficients, but standard errors, and associated inference procedures, could still be affected.

[6]Crafts and Toniolo (1996) comment that 'Houthakker and Magee's results are now generally regarded as being unreliable.' (p. 12).

[7]The full cointegration theory is more general, referring to $\mathsf{I}(d)$ processes, but the $\mathsf{I}(1)$ case is most relevant to the applications discussed here.

form; then,

$$\boldsymbol{\beta}'\mathbf{z}_t = \beta_1 \mathsf{m}_t + \beta_2 \mathsf{rpm}_t + \beta_3 \mathsf{yd}_t \sim \mathsf{I}(0)$$

where $\mathbf{z}_t = (\mathsf{m}_t, \mathsf{rpm}_t, \mathsf{yd}_t)'$ and $\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3)'$ represents the long-run relationship vector.

One popular way to use this theory has been to incorporate potential cointegrating relationships within a vector equilibrium-correction (VECM) model, which has two principal attractions. First, it allows for potential endogeneity in the variables in a trade system, and secondly it is sufficiently flexible to allow for more than one cointegrating vector. In this specification, the long-run relationship between each trade flow and its price and income drivers can be captured in terms of levels, whilst the short-run elasticities are given by the response of trade growth to changes in price and income.

As a first step, a unit-root test should be used to check that the variables are $\mathsf{I}(1)$.[8] To do this, a conventional approach in time-series studies such as Hooper, Johnson and Marquez (2000) or Fawcett (2003) is to perform an Augmented Dickey-Fuller test on the levels of the variables, and their differences, to determine their order of integration. Applying this test to UK data from 1974 to 2000 across manufacturing and service sectors, Fawcett (2003) finds that all of the variables of interest in (2.1) are $\mathsf{I}(1)$, justifying the concern raised above regarding the use of techniques not suited to non-stationary data, and suggesting that a VECM could be an appropriate model to use. Similarly, Hooper *et al.* (2000) find evidence of $\mathsf{I}(1)$ data in a study of aggregate (whole-economy) data for the G7 countries.

These existing studies use a VECM comprising a trade flow variable, relative price measure, and income term. Thus, for example, the system for

[8] Of course, although a unit root is a sufficient condition for non-stationarity, it is not necessary. However, as an earlier point notes, the VECM approach for the case of $\mathsf{I}(d)$ processes is of greatest interest here.

imports is expressed as:[9]

$$\Delta \mathbf{z}_{mt} = \boldsymbol{\kappa}_m + \sum_{j=1}^{n} \boldsymbol{\Gamma}_{mj} \Delta \mathbf{z}_{m,t-j} + \boldsymbol{\Pi}_m \mathbf{z}_{m,t-1} + \boldsymbol{\epsilon}_{mt} \tag{2.2}$$

where $\boldsymbol{\epsilon}_{mt} \sim \mathsf{IN}(0, \boldsymbol{\Omega}_m)$, $\mathbf{z}_{mt} = (\mathsf{m}_t, \mathsf{rpm}_t, \mathsf{yd}_t)'$ (lower case denoting logs of variables), $\boldsymbol{\kappa}_m$ is an intercept vector (possibly constrained to lie in the cointegration space), $\boldsymbol{\Gamma}_{mj}$ are the short-run feedback coefficients and $\boldsymbol{\Pi}_m$ combines the cointegrating combinations and their loadings, such that $\boldsymbol{\Pi}_m = \boldsymbol{\alpha}_m \boldsymbol{\beta}_m'$. Thus in a cointegrated system, $\boldsymbol{\alpha}_m, \boldsymbol{\beta}_m \in \mathbb{R}^{3\times r}$ where $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are of full rank $0 < r < 3$, where r represents the number of cointegrating relationships, and $\boldsymbol{\Pi}_m$ has reduced rank $(3-r)$. The corresponding equation for exports comprises $\mathbf{z}_{xt} = (\mathsf{x}_t, \mathsf{rpx}_t, \mathsf{yf}_t)'$ with analogues $\boldsymbol{\epsilon}_{xt}$, $\boldsymbol{\kappa}_x$, $\boldsymbol{\Gamma}_x$ and $\boldsymbol{\Pi}_x$ elements. In this framework, the parameters of interest are contained in the $\boldsymbol{\beta}'$ matrix, as shown in the example above (where subscripts have been omitted for notational simplicity), which are the long-run relationships between the variables in $\mathbf{z}_t$. Using (2.2) and its export equivalent, Hooper *et al.* (2000) derive estimates for trade elasticities for the G7 countries based on aggregate data across several decades (depending on country, for the mid-1950s/70s to 1994), and these are presented in Tables 2.1 and 2.2.

First testing their model to establish the rank of $\boldsymbol{\Pi}_x$ and $\boldsymbol{\Pi}_m$ using Johansen's trace and maximised eigenvalue tests, they find evidence of one cointegrating vector (i.e. $r = 1$) across all seven countries, after a suitable choice of lags in the feedback term, which is given by n in (2.2). Then estimation of this cointegrating vector follows via maximum likelihood estimation of a reduced rank regression (see Johansen (1995)) to yield estimates of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, and from there it is possible to test hypotheses on the parameters.

Following the determination of long-run relationships, the short-run elasticities can be obtained by modeling the variables in a more parsimonious $\mathsf{I}(0)$ setting. If the system aspect from above is preserved, then one way to do

[9] This corresponds to equation (2) of Hooper *et al.* (2000), with the export equation given by their equation (1). Fawcett (2003) uses their framework for his estimation exercise, and the statistical theory underlying the VECM approach used, is due to Johansen (1995).

Table 2.1: TRADE ELASTICITIES IN THE LITERATURE: IMPORTS

		HJM		HM		Krugman	
		Price	Income	Price	Income	Price	Income
Belgium	*LR*			-1.02^{*}	1.94^{*}	-0.53^{*}	1.99^{*}
	SR						
Canada	*LR*	-0.9^{*}	1.4^{*}	-1.46^{*}	1.20^{*}	-1.45	1.66^{*}
	SR	-0.1	1.3^{*}				
Denmark	*LR*			-1.66^{*}	1.31^{*}		
	SR						
France	*LR*	-0.4^{*}	1.6^{*}	0.17	1.66^{*}		
	SR	-0.1	1.7^{*}				
Germany	*LR*	-0.06^{*}	1.5^{*}	$-0.24^{\dagger}$	$1.80^{*\dagger}$	$-0.09^{\dagger}$	$2.83^{*\dagger}$
	SR	-0.2^{*}	1.0^{*}				
Italy	*LR*	-0.4^{*}	1.4^{*}	-0.13	2.19^{*}	-0.68^{*}	3.65^{*}
	SR	-0.0	1.0^{*}				
Japan	*LR*	-0.3^{*}	0.9^{*}	-0.72^{*}	1.23^{*}	-0.42	0.80
	SR	-0.1	1.0^{*}				
Netherlands	*LR*			0.23	1.89^{*}	-0.22	2.66^{*}
	SR						
Norway	*LR*			-0.78^{*}	1.40^{*}		
	SR						
UK	*LR*	-0.6	2.2^{*}	0.22	1.66^{*}	0.99^{*}	-0.20^{*}
	SR	-0.0	1.0^{*}				
US	*LR*	-0.3^{*}	1.8^{*}	-0.54	1.51^{*}	-0.93^{*}	1.31^{*}
	SR	-0.6	2.3^{*}				
Estimation method		VECM-MLE (LR), OLS (SR)		OLS in levels		OLS in levels	
Data span and frequency		Quarterly, 1950/70-1994		Annual, 1951-1966		Annual, 1971-1986	

NOTES: LR and SR relate to estimates of long-run and short-run elasticities, where papers distinguish between the two; otherwise it is assumed that papers report long-run parameter estimates. An * denotes statistical significance at the 5% level. Data for Germany denoted † relate to West German estimates. HJM represents Hooper, Johnson and Marquez (2000); HM is Houthakker and Magee (1969); Krugman is his (1989) paper.

Table 2.2: TRADE ELASTICITIES IN THE LITERATURE: EXPORTS

		HJM		HM		Krugman	
		Price	Income	Price	Income	Price	Income
Belgium	*LR*			0.42	1.83*	−0.19*	1.24*
	SR						
Canada	*LR*	−0.9*	1.1*	−0.59*	1.41*	0.8*	2.87*
	SR	−0.5*	1.1*				
Denmark	*LR*			−0.56	1.69*		
	SR						
France	*LR*	−0.2	1.5*	−2.27*	1.53*		
	SR	−0.1	1.8*				
Germany	*LR*	−0.3	1.4*	1.70†	2.08*†	−0.55†	2.15*†
	SR	−0.1	0.5				
Italy	*LR*	−0.9*	1.6*	−0.03	2.95*	−0.23	2.41*
	SR	−0.3*	2.3*				
Japan	*LR*	−1.0*	1.1*	−0.80	3.55*	−0.88*	1.65*
	SR	−0.5*	0.6				
Netherlands	*LR*			−0.82	1.88*	−0.76	3.86*
	SR						
Norway	*LR*			0.20	1.59*		
	SR						
UK	*LR*	−1.6*	1.1*	−0.44	0.86*	−0.54	1.30*
	SR	−0.2*	1.1*				
US	*LR*	−1.5*	0.8*	−1.51*	0.99*	−1.42*	1.70*
	SR	−0.5*	1.8*				
Estimation method		VECM-MLE (LR), OLS (SR)		OLS in levels		OLS in levels	
Data span and frequency		Quarterly, 1950/70-1994		Annual, 1951-1966		Annual, 1971-1986	

NOTES: See notes for Table 2.1.

this is via FIML estimation of a parsimonious VAR (PVAR) as discussed by Hendry (1995, chapter 16), in the context of a system examining real money demand. An alternative method is pursued by Hooper *et al.* (2000), who explore a single-equation equilibrium-correction model which, for imports,

is given by:[10]

$$\begin{aligned}\Delta \mathsf{m}_t =& \mu_m + \sum_{j=1} \pi_{mj} \Delta \mathsf{m}_{t-j} + \sum_{j=0} \tau_{mj} \Delta \mathsf{rpm}_{t-j} + \sum_{j=0} \rho_{mj} \mathsf{yd}_{t-j} \\ &+ \mathbf{\Phi}_m \widehat{ECM}_{m,t-1} + \varepsilon_{mt} \end{aligned} \tag{2.3}$$

where $\varepsilon_{mt} \sim \mathsf{IN}\left(0, \sigma_m^2\right)$, and $\widehat{ECM}_{m,t-1} = \widehat{\boldsymbol{\beta}}' \mathbf{z}_{t-1}$, the estimated disequilibrium component from the Johansen estimation above. In this formulation, the short-run coefficients on price (τ_m) and income (ρ_m) are normalised by the sum of coefficients on the lagged dependent variable, so that

$$\tau_m = \frac{\sum_j \tau_{mj}}{1 - \sum_j \pi_{mj}} \quad \text{and} \quad \rho_m = \frac{\sum_j \rho_{mj}}{1 - \sum_j \pi_{mj}}$$

and they measure the cumulative impact on import growth of recent price and income changes. The adjustment back to long-run equilibrium is captured by $\frac{\Phi_m}{1-\sum_j \pi_{mj}}$.

This general approach has two main advantages over the earlier methodology. First, it provides a systematic treatment of non-stationarity in the data, which itself was revealed via unit root tests; and secondly, the distinction between long- and short-run dynamics may be useful, since it disentangles, to an extent, the immediate impact of changes in the drivers of trade, and any long-term influences which may be driven by different factors. This sits well with the theoretical points discussed above, which point to a more microeconomic supply-side role for the determination of long-run trade performance, whilst macroeconomic policy affects the economy in the short-run, a perspective that could be relevant from a policy-making point of view, for example.

However, problems still remain, as the existing literature has largely focused on an aggregate picture, which not only precludes an assessment of trade behaviour by sector, but could also affect the validity of any elasticity estimates obtained, in a reasonably wide range of situations. Thus the following section introduces the main empirical work of this chapter, which

[10] This differs from the import equation in the I(0) PVAR since it conditions on current values of the other variables.

seeks to estimate disaggregated elasticities for a number of sectors in a panel of countries.

2.2 Building a panel model to estimate trade elasticities

The choice of a panel dataset, and associated estimation techniques, is motivated by two considerations:

- First, it can be used to obtain sector-by-sector trade elasticities even when the number of time periods available is quite small, by exploiting similarities in industries across countries (thereby holding an advantage over a conventional time-series approach using disaggregated data), and;
- Secondly, even if the focus is purely on an economy-wide picture of trade performance, panel data methods applied to disaggregated data can offer more reliable estimates in a wide range of situations.

More concretely, there are a number of advantages of a panel approach.

First, with regard to data on trade flows, relative prices, and income, it is quite likely that aggregate data disguises significant heterogeneity in the same series, when they are compared on a sectoral level. To see this more clearly, Figure 2.1 compares the relative price of exports from two US manufacturing sectors over the same period of time. The remarkable difference in behaviour across each sector is explained by differences in trade flows, in addition to conventional differences in the actual producer price of output. Thus whilst over 95% of US exports of wood and wood products were destined for Canada, only approximately 15% of electrical and optical equipment (including IT) were, with Japan instead being the major destination (receiving 65% of exports).[11] Over the period shown, the US dollar consistently appreciated against the Canadian dollar, whilst the bilateral exchange rate with the Yen was more volatile, with a sharp depreciation over the first five years of the sample, followed by an appreciation. Thus even

[11] Based on 1995 trade figures.

Figure 2.1: HETEROGENEITY IN US RELATIVE EXPORT PRICES

SOURCE: Author's calculations, using data for two manufacturing sectors in the US, from the OECD Structural Analysis (STAN) database (ISIC codes provided in Appendix 2.A.1). Log relative export prices are calculated by a trade-weighted export price series for each sector, divided by the price of domestic output from the same sector.

in this simple comparison it is evident that there is important time-series variation in the disaggregated explanatory variables, and this may be lost by averaging in the disaggregated series. As a result, an aggregate estimate of, say, US trade elasticities, might be unbiased in an econometric sense as a measure of 'average' economy-wide trade behaviour, but it could disguise significant variation in the response across a more disaggregated picture. Of course, there are more than two sectors in the US economy, and so a more systematic treatment of cross-sector variation in the variables of interest is presented below, for a number of countries.

Besides differences in the *variables*, an important source of variability lies in the *parameters* themselves. It could well be the case that the trade *elasticities* for wood and IT sectors differ, and this would not, of course, be detected in an aggregate study. In the presence of parameter heterogeneity across sectors, though, there could be problems if aggregated data are used to obtain a set of 'average' trade elasticities. Pesaran and Smith (1995) show that in dynamic models, using an aggregating procedure can give inconsistent parameter estimates, even if the number of sectors and time periods

is large. To see why, consider a simple heterogeneous dynamic model such as:[12]

$$y_{it} = \alpha_i y_{i,t-1} + \boldsymbol{\beta}_i' \mathbf{x}_{it} + \epsilon_{it} \qquad \begin{array}{l} i = 1, 2, \ldots, N \\ t = 1, 2, \ldots, T \\ \epsilon_{it} \sim \mathsf{IN}\left(0, \sigma_i^2\right) \end{array} \tag{2.4}$$

for some $\mathsf{I}(0)$ stochastic $(y, \mathbf{x}')'$, with coefficients α_i and $\boldsymbol{\beta}_i$ given by a random coefficients model so that

$$\alpha_i = \alpha + \eta_{1i} \tag{2.5a}$$

$$\boldsymbol{\beta}_i = \boldsymbol{\beta} + \boldsymbol{\eta}_{2i} \tag{2.5b}$$

where η_{1i} and $\boldsymbol{\eta}_{2i}$ have zero means and constant covariances. Then, Pesaran and Smith show that using the group (i.e. cross-section) averages of y_{it} and $\mathbf{x}_{it}$, denoted $\overline{y}_t$ and $\overline{\mathbf{x}}_t$, the aggregate time-series regression is given by

$$\overline{y}_t = \alpha \overline{y}_{t-1} + \boldsymbol{\beta}' \overline{\mathbf{x}}_t + \overline{v}_t \tag{2.6}$$

where

$$\overline{v}_t = \overline{\epsilon}_t + \frac{1}{N} \sum_{i=1}^{N} (\eta_{1i} y_{i,t-1} + \boldsymbol{\eta}_{2i}' \mathbf{x}_{it}).$$

Thus estimation of (2.6) via OLS will yield inconsistent estimates if the $\mathbf{x}_{it}$ are serially correlated, and traditional responses to this via instrumental variables estimation will also fail, since anything correlated with lagged values of $\mathbf{x}_{it}$ will also be correlated with $\overline{v}_t$.[13] Further, since the variables in each trade equation are measured in log form, aggregation bias can emerge since the sum of logged terms is not equal to the log of aggregated series.

As later sections explore, alternative panel techniques can produce more reliable estimates of the $\boldsymbol{\beta}_i$, and examine the possibility that $\boldsymbol{\beta}_i = \boldsymbol{\beta}$ for all i (i.e. the case of complete homogeneity) amongst other cases. Further, the panel dimension is attractive even if there is homogeneity across groups in the $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ parameters, or at least, homogeneity in subsets of groups, as it offers an avenue away from the problem of low statistical power in small samples that affects the Johansen cointegration estimator. If some

[12] This draws on Pesaran and Smith (1995) equations (2.1), (2.2) and (2.12).

[13] See Pesaran and Smith (1995, p. 86).

degree of parameter homogeneity is assumed, then meaningful estimates of long-run relationships can be obtained even for relatively small time-dimension datasets, provided that the number of individuals in the panel is large enough.[14]

This last point suggests an interesting role for two different cross-section dimensions in the empirical problem discussed here. On one hand, the foregoing discussion has considered a disaggregated view within a country, in which total output is decomposed into its sectoral origins, for example. However, there is another dimension available, which looks at the same industry – such as wood and wood products – across several countries. Examining *several* sectors across *multiple* countries raises the possibility of an N-S-T panel where N represents the number of countries, S the number of sectors, and T the time dimension. Although conventionally this might be treated as an $NS \times T$-dimension panel in which all country-industry combinations are separate individuals, the time span of the dataset here is too short for estimators or procedures that require a large number of time observations. Thus out of pragmatic considerations, there is an appeal in exploiting the N/S distinction here. For example, when making assumptions about the degree of homogeneity across individuals in the panel, there could be an argument for pooling data across countries, and thereby allowing heterogeneity across industries. The rationale for this is that similarities in the behaviour of each industry across countries – due to common technology and product characteristics, say – would suggest that there is a greater likelihood of elasticities being common to industries, rather than countries. Thus for example, whilst it may be plausible that the elasticities for textiles are the same in the US and the UK, it might not be so for US textiles compared to US IT.

[14] In this sense, individuals refers to distinct groups, whether countries, sectors, firms etc.

2.2.1 The panel dataset

The data used in this chapter comes primarily from the OECD Structural Analysis database, known as STAN (OECD 2005),[15] and covers the period 1988 – 2002. STAN includes data on nearly 50 separate industry groupings (including services), across all OECD member countries. However, as Appendix 2.A.1 explains, the coverage across countries and industries is variable, so that a subset of the data is studied here, covering 13 sectors in 11 countries in a balanced panel for all 15 years. This leaves an N-S-T data 'cube' of nearly equal proportions on each side, and has implications for the choice of panel estimator, since not all behave in the same way for given individual and time dimensions.

2.2.2 Approaches to dynamic panel estimation

The role of heterogeneity in industry and country elasticities can be shown explicitly in two 'stylized' trade equations:

$$\mathsf{m}_{ist} = \beta_{mis}\mathsf{rpm}_{ist} + \gamma_{mis}\mathsf{yd}_{ist} \tag{2.7}$$

$$\mathsf{x}_{ist} = \beta_{xis}\mathsf{rpx}_{ist} + \gamma_{xis}\mathsf{yf}_{ist} \tag{2.8}$$

$$i = 1, \ldots, N; \quad s = 1, \ldots, S; \quad t = 1, \ldots, T$$

using the same variable definitions as equation 2.1. Clearly these do not represent regression equations that one might estimate directly, as they miss obvious elements (such as error specifications) and do not specify the nature of any dynamics. However, (2.7) and (2.8) are useful in that they show how each country and sector pairing can be treated, and how this relates to the question of heterogeneity in parameters, and the dependent and independent variables.

The *is* subscript on all three variables in each equation demonstrates that they can all be country and sector specific in this framework. As Figure 2.1 shows in the case of rpx for the US wood and electrical equipment industries, this is a valuable advantage of the model, over a more aggregated case.

[15] Available from http://www.oecd.org/sti/stan.

Another point is that assumptions about the behaviour of the parameters β and γ can be varied. Based on the discussion above with regard to the extent to which the data can be pooled, there are four main hypotheses that can be considered in an empirical modelling strategy:[16]

$$\begin{array}{lll}
\mathsf{H}_{\text{HET}}: & \text{Complete Heterogeneity} & \beta_{is} = \beta_{is} \\
 & & \gamma_{is} = \gamma_{is} \\
\mathsf{H}_{\text{NOHET}}: & \text{No Heterogeneity} & \left.\begin{array}{l}\beta_{is} = \beta \\ \gamma_{is} = \gamma\end{array}\right\} \text{ for all } i, s \\
\mathsf{H}_{\text{I}}: & \begin{array}{l}\text{Industry Heterogeneity} \\ \text{(Country homogeneity)}\end{array} & \left.\begin{array}{l}\beta_{is} = \beta_{s} \\ \gamma_{is} = \gamma_{s}\end{array}\right\} \text{ for all } i \\
\mathsf{H}_{\text{C}}: & \begin{array}{l}\text{Country Heterogeneity} \\ \text{(Industry homogeneity)}\end{array} & \left.\begin{array}{l}\beta_{is} = \beta_{i} \\ \gamma_{is} = \gamma_{i}\end{array}\right\} \text{ for all } s
\end{array}$$

A number of observations follow from these four cases. H_{HET} essentially treats the data on country-industry pairs as separate time series, so there are NS estimates of β and γ, each corresponding to the relevant is individual. In contrast, $\mathsf{H}_{\text{NOHET}}$ assumes the opposite, allowing the data to be treated as a panel with NS individuals, yielding one estimate of β and γ from the panel dataset. The intermediate cases H_{I} and H_{C} are a feature of the N/S distinction in the cross-section dimension of the data, and offer an attractive compromise in the choice between assuming complete heterogeneity (and thereby losing some of the benefits of the panel approach) and making the potentially restrictive assumption of complete homogeneity. In either intermediate case, data are pooled across countries (for H_{I}) or industries (for H_{C}), to get either S estimates of β and γ in the former case, or N country-specific coefficients in the latter. As observed above, if a homogeneity assumption must be imposed (due to the poor performance of estimators or tests in small-T samples, say) then there could be an argument in favour of hypothesis H_{I}, on the grounds that there might be significant similarity in the same industry across countries. However, Tables 2.1 and 2.2 both re-

[16] In a full specification with dynamics and error terms these restrictions would apply to all parameters, of course, but for now these four scenarios convey the central point.

port that on an aggregate level, at least, there are significant cross-country differences, and it might be dangerous to ignore these *ex ante*.

Thus the estimation strategy pursued here follows a hybrid approach, by exploiting the distinction between the long-run cointegrating relationship that exists for each country-sector individual unit, and short-run changes in the variables in (2.8) and (2.7). Imposing an intermediate level of homogeneity in the long-run relationship and allowing for different degrees of heterogeneity in the short-run responses offers a suitable compromise to the trade-off set out above.

In order to understand how such a hybrid approach fits in the literature on dynamic panel models with possibly heterogeneous coefficients, a starting point is the analysis of Pesaran and Smith (1995), which was discussed above. In their baseline model, given earlier in equations (2.4) and (2.5), they study a single-equation dynamic model with additional explanatory variables (given by $\mathbf{x}_{it}$), where the slope coefficients on the lagged dependent variable and the $\mathbf{x}_{it}$s, denoted α_i and $\boldsymbol{\beta}_i$, differ randomly across individuals i. In this setting, Pesaran and Smith show that some common panel estimators, including the pooling approach, and aggregating procedure outlined above, deliver biased estimates if there is genuine parameter heterogeneity and the x_{it}s are serially correlated.[17] The underlying cause of this bias is serial correlation in the disturbance, which is evident in (2.6) when using aggregate data; in the pooling case the model is given by:[18]

$$y_{it} = \delta_i + \alpha y_{i,t-1} + \boldsymbol{\beta}' \mathbf{x}_{it} + v_{it}$$

where

$$v_{it} = \epsilon_{it} + \eta_{1i} y_{i,t-1} + \boldsymbol{\eta}_{2i}' \mathbf{x}_{it}.$$

As Pesaran and Smith show, for $\mathsf{I}(0)$ series the correlation between the x_{it} and v_{it} render the usual pooled estimators of α and β inconsistent, even for large T and/or N. Thus the choice of pooling assumption $\mathsf{H}_{\text{NOHET}}$, H_{I}

[17] These concerns would almost certainly apply to the trade equations here.

[18] These equations correspond to Pesaran and Smith (1995) equations (2.4) and (2.5).

or H_{C} could be all the more important, since an incorrect choice introduces *econometric* problems, in addition to the *economic* judgment over which assumption is more plausible.

In the specification given by (2.4), Pesaran and Smith consider the 'long-run' effect of $\mathbf{x}$ on y, defining $\overline{\boldsymbol{\theta}}$ to be the appropriate measure of the average, where

$$\boldsymbol{\theta}_i = \frac{\boldsymbol{\beta}_i}{1-\alpha_i}; \quad \overline{\theta} = \frac{1}{N}\sum_{i=1}^{N}\theta_i; \quad i = 1, \ldots, N$$

If y and $\mathbf{x}$ are cointegrated, then this long-run relationship corresponds to a cointegrating vector (assuming that $r = 1$) equal to $(1, -\boldsymbol{\theta}_i)'$. As an alternative to the aggregating and pooling methods, they also consider running N separate time-series regressions of (2.4) for $i = 1, \ldots, N$ and then taking the average of the estimated coefficients (this is known as the Mean Group estimator). They also evaluate the cross-section approach using time-averages of all the variables, and then running a cross-section regression. For very short time spans, small-sample bias might render the latter approach unreliable, but in a parsimonious model there could be scope for using the Mean Group method to get an impression of the distribution of parameter estimates, before comparing them to results from pooled alternatives.

In the context of the hybrid homogeneity assumption, Pesaran, Shin and Smith (1999) distinguish between heterogeneity assumptions in the short-run and long-run coefficients in a model. Their Pooled Mean Group approach allows α_i and $\boldsymbol{\beta}_i$ to be heterogeneous across individuals, whilst it restricts $\boldsymbol{\theta}$ to be common. In a cointegrated $\mathsf{I}(1)$ setting, this is analogous to allowing short-run coefficients and equilibrium-correction loadings to be heterogeneous, whilst imposing a common long-run cointegrating vector. Their rationale for this is that "budget or solvency constraints, arbitrage conditions or common technologies" may influence all groups in similar ways in the long run, but in the short run it is less obvious that this will be the case.[19] This distinction raises the question of how to obtain estimates of long- and short-run coefficients in a panel context. The earlier discussion on

[19] Pesaran *et al.* (1999, p. 621).

the existing trade literature showed that modern time series-based studies use cointegration theory as the statistical backdrop for estimation, and so it would seem reasonable the try and implement this in the panel.

2.2.2.1 Long-run estimation in cointegrated panel models

How do time-series considerations of nonstationarity and cointegration translate to the panel setting? In several respects the two literatures share common concerns;[20] first, the variables included in a model must be tested for their order of integration, and then estimation and testing of any cointegrating vectors takes place, either in a single-equation framework, or in a system. However, with the panel dimension comes a number of new issues to resolve, revolving around some key themes. The effects of parameter heterogeneity that were discussed above, are also relevant here; and there is a further concern relating to the interpretation of test results, such that even only slight heterogeneity (in the parameters in a unit-root test, for instance) can lead to incorrect rejection of hypotheses during testing. Secondly, the dimension of the panel is important, as some approaches are best suited to large T datasets, whilst others work with small T but large N. Finally, an important assumption made in some procedures is that the individuals in the panel are independent, so that a shock in one will not affect another; clearly in the case of country/sector data this is a difficult assumption to justify, and so methods robust to some form of cross-section dependence will be necessary.

A simple model considered by Pesaran *et al.* (1999) can illustrate how some of these factors come into play, and leads on to alternative estimation strategies. This assumes that the variables considered are actually I(1), and that there is a cointegrating relationship between them; testing that this is true is, of course, vital, and so this section finishes by examining some relevant panel unit root tests.

Using a reparameterized ARDL $(p, q, q, \ldots, q)$ model, the data generating

[20] A comprehensive overview of panel time series is provided by Smith and Fuertes (2007), whilst unit roots and cointegration in panels are specifically considered by Breitung and Pesaran (2008).

process that Pesaran *et al.* (1999) use can be written as:[21]

$$\Delta y_{it} = \phi_i y_{i,t-1} + \boldsymbol{\beta}_i' \mathbf{x}_{it} + \sum_{j=1}^{p-1} \lambda_{ij}^* \Delta y_{i,t-j} + \sum_{j=0}^{q-1} \boldsymbol{\delta}_{ij}^{*\prime} \Delta \mathbf{x}_{i,t-j} + \mu_i + \epsilon_{it} \quad (2.9)$$
$$i = 1, 2, \ldots, N \quad t = 1, 2, \ldots, T$$

where y_{it} is the dependent variable, $\mathbf{x}_{it}$ is a $k \times 1$ vector of explanatory variables, μ_i are individual-specific fixed effects, ϕ_i, $\boldsymbol{\beta}_i$, λ_{ij}^* and $\boldsymbol{\delta}_{ij}^{*\prime}$ are parameters which in turn are functions of the parameters in the benchmark ARDL representation (see equation (1) in Pesaran *et al.* (1999)). This corresponds to a single-equation equilibrium-correction representation in the case where there is only one cointegrating vector. If there are enough time periods to run separate Pooled Mean Group regressions for each i, they then stack the time-series observations and then derive a maximum likelihood estimator for the parameters of interest, based on several assumptions, of which three are of particular note here. First, they assume that the error ϵ_{it} in (2.9) is independently distributed across i, with mean zero and variance σ_i^2. Secondly, the underlying roots of the ARDL model are stable, such that $\phi_i < 0$, ensuring that there is a long-run relationship between y_{it} and $\mathbf{x}_{it}$ given by $y_{it} = (\boldsymbol{\beta}_i'/\phi_i)\mathbf{x}_{it} + \eta_{it}$ for each i, with η_{it} being a stationary error. Finally, defining $\boldsymbol{\theta}_i = (\boldsymbol{\beta}_i'/\phi_i)$, they assume homogeneity of long-run coefficients across groups, such that $\boldsymbol{\theta}_i = \boldsymbol{\theta}$ for all i.

With a further assumption that the errors ϵ_{it} are normally distributed, the joint concentrated log-likelihood for all individuals over time can be derived by taking the product of each individual i log-likelihood, which highlights the importance of the independence assumption, and also identifies a major problem with using the approach in this chapter. Since the independence assumption is almost certainly violated here, due to country-specific shocks affecting all industries, for instance, the joint likelihood cannot be derived as the product of the individual likelihoods. Further, given that the independence violation occurs over *individuals* rather than across *time*, and

[21] This corresponds to their equation (2).

the exact nature of dependence is unknown, a possible route of taking the conditional model to construct the joint likelihood is unavailable.

For similar reasons, other panel estimators such as the Fully-Modified (FMOLS) method are likely to be inappropriate. The FMOLS estimator (see Pedroni (1995, 2000) or Phillips and Moon (1999)) corrects for long-run endogeneity in the regressors, and allows for heterogeneous short-run dynamics. However, since Pedroni (2000), Assumption 1.2 (p. 99) requires cross-sectional independence, it is unsuitable for the kind of data used here. Further, some correction mechanisms that Pedroni (2000) suggests, such as estimating a full set of cross-section dependencies and using them to pre-multiply the errors, require significantly larger T dimension than is available here. Even on the level of the convergence theorems that underlie the asymptotic properties of the FMOLS estimator, Breitung (2005) points out that "Phillips and Moon (1999, p. 1092) state that '... when there are strong correlations in a cross-section (as there will be in the face of global shocks) we can expect failures in the strong laws and central limit theory arising from the nonergodicity' " (p. 160).

Whilst contemporaneous correlation causes problems for some estimators, there are alternatives that are robust to it, and of these, three are prominent, all of which allow for multiple cointegrating vectors, in contrast to earlier techniques which assumed only one long-run relationship. The first two seek to extend the time-series VECM model of Johansen (1995) into a panel dimension, building on a likelihood-based framework for estimation. Thus both Larsson and Lyhagen (1999) and Groen and Kleibergen (2003) start with an underlying panel VECM representation, which is based on a VAR(m) model for a p-dimensional random vector $\mathbf{Y}_{it}$, where for individuals $i = 1, \ldots, N$, there are $0 < r_i < p$ cointegrating relationships. Stacking the $\mathbf{Y}_{it}$ vectors across all individuals, yields a familiar representation in the form of:

$$\Delta \mathbf{Y}_t = \mathbf{A}\mathbf{B}'\mathbf{Y}_{t-1} + \sum_{k=1}^{m-1} \mathbf{\Gamma}_k \Delta \mathbf{Y}_{t-k} + \boldsymbol{\epsilon}_{it} \tag{2.10}$$

where $\mathbf{A}$ is an $Np \times \sum_i r_i$ matrix of short-run coefficients, $\mathbf{B}$ is an $Np \times \sum_i r_i$

matrix of long-run cointegrating vectors, the $\mathbf{\Gamma}_k$ are a set (for $k = 1, \ldots, m-1$) of feedback coefficients, and the error $\boldsymbol{\epsilon}_t$ has a multivariate normal density given by $\boldsymbol{\epsilon}_t \sim \mathsf{N}(\mathbf{0}, \mathbf{\Omega})$, where $\mathbf{\Omega}$ is an $Np \times Np$ covariance matrix with, crucially, an off-diagonal structure.

In the face of this rather compact representation of the models both papers use, it is perhaps helpful to highlight their important features. First, whilst both models allow for multiple cointegrating vectors, in the Groen and Kleibergen (2003) case, it is assumed that $r_i = r$ for all i, such that all individuals in the panel share the *same* set of cointegrating vectors $\mathbf{B}$.[22] Further, in both cases it is assumed that off-diagonal elements in the block $\mathbf{B}$ are zero, ruling out between-group cointegration.[23] The papers differ in their treatment of the short-run error correction parameters in $\mathbf{A}$, as Larsson and Lyhagen (1999) allow for non-zero off-diagonal elements, whilst Groen and Kleibergen (2003) assumes a diagonal structure. Similarly, the former allow for a non-diagonal set of $\mathbf{\Gamma}_k$ matrices, whilst the latter does not. Perhaps most importantly, given the context of this discussion, is the fact that *both* approaches allow for a non-diagonal error covariance matrix $\mathbf{\Omega}$, which means that contemporaneous correlation is allowed in the model, thereby allaying one of the chief concerns of the earlier estimation methods.

In order to derive estimates of the long-run $\mathbf{B}$ parameters, the approach of Larsson and Lyhagen (1999) is to obtain a preliminary estimate of the diagonal elements of $\mathbf{B}$ (by running N standard time-series cointegrating regressions), and then using these to estimate $\mathbf{A}$ and $\mathbf{\Omega}$, switching back and forth from updated estimates of $\mathbf{B}$, and $\mathbf{A}$ and $\mathbf{\Omega}$ until there is no further increase in the likelihood. In contrast, Groen and Kleibergen (2003) show that there is a relationship between the maximum likelihood estimator, which has an unknown analytical expression in the panel setting, and a GMM estimator, and then use iterated GMM estimation to obtain maximum likelihood estimates for the coefficients.

[22] Larsson and Lyhagen (1999) also discuss a variant of their own model in which homogeneity in cointegrating vectors is imposed.

[23] Banerjee, Marcellino and Osbat (2004) emphasise some of the problems that can arise when between-group cointegration is ignored.

Whilst both of these approaches are attractive, in that they allow for multiple cointegrating vectors, and contemporaneous correlation in disturbances across individuals, they suffer from the need for a relatively large number of time observations. Thus for example, in order to estimate (2.10) for a three-variable trade model across all 143 individuals (using assumption $\mathsf{H}_{\text{NOHET}}$), Larsson and Lyhagen (1999) require at least $T = NSp + 2 = 431$ time periods, whilst the asymptotic results of Groen and Kleibergen (2003) rely on large T. In the context of our panel dataset, where $T = 15$, it is clear that such requirements will not be satisfied.

In response to the small-sample limitations of the these models, Breitung (2005) proposes an alternative method that is possible to implement even with a sample size of 15, and that also allows for contemporaneous correlation in disturbances. In a simple cointegrated VAR(1) model such as:

$$\Delta \mathbf{y}_{it} = \boldsymbol{\alpha}_i \boldsymbol{\beta}' \mathbf{y}_{i,t-1} + \boldsymbol{\epsilon}_{it} \tag{2.11}$$
$$i = 1, \ldots, N \quad t = 1, \ldots, T \quad \mathsf{E}[\boldsymbol{\epsilon}_{it}] = \mathbf{0} \quad \boldsymbol{\Sigma}_i = \mathsf{E}[\boldsymbol{\epsilon}_{it}\boldsymbol{\epsilon}_{it}']$$

his approach, which seeks to estimate the long-run cointegrating vector $\boldsymbol{\beta}$ (which is common to all i), splits the estimation process into two steps. First, consistent estimates of $\boldsymbol{\alpha}_i$ and $\boldsymbol{\Sigma}_i$ are obtained (denoted $\widehat{\boldsymbol{\alpha}}_i$ and $\widehat{\boldsymbol{\Sigma}}_i$), either via the Johansen (1995) maximum likelihood estimator in the case of multiple cointegrating vectors, or the Engle-Granger two-step method if there is only one such vector.[24] Then, Breitung shows that the $\boldsymbol{\beta}$ matrix can be estimated by a pooled OLS regression of:[25]

$$\widehat{\mathbf{z}}_{it} = \boldsymbol{\beta}' \mathbf{y}_{i,t-1} + \widehat{\mathbf{v}}_{it}$$
$$i = 1, \ldots, N \quad t = 1, \ldots, T$$

where $\widehat{\mathbf{z}}_{it} = (\widehat{\boldsymbol{\alpha}}_i' \widehat{\boldsymbol{\Sigma}}_i^{-1} \widehat{\boldsymbol{\alpha}}_i)^{-1} \widehat{\boldsymbol{\alpha}}_i' \widehat{\boldsymbol{\Sigma}}_i^{-1} \Delta \mathbf{y}_{it}$ and $\widehat{\mathbf{v}}_{it} = (\widehat{\boldsymbol{\alpha}}_i' \widehat{\boldsymbol{\Sigma}}_i^{-1} \widehat{\boldsymbol{\alpha}}_i)^{-1} \widehat{\boldsymbol{\alpha}}_i' \widehat{\boldsymbol{\Sigma}}_i^{-1} \boldsymbol{\epsilon}_{it}$. By normalising the cointegrating vectors such that $\boldsymbol{\beta} = [\mathbf{I}_r, \mathbf{B}]'$,[26] and dividing

[24] These estimates are consistent as $T \to \infty$.

[25] This corresponds to Breitung (2005) equation (4), p. 156.

[26] Note that this $\mathbf{B}$ is not the same as $\mathbf{B}$ in equation 2.10. As Breitung and Pesaran (2008) note, other exact-identifying restrictions can be used without affecting the result.

the vector $\mathbf{y}_{it}$ into two subvectors, $\mathbf{y}_{it}^{(1)}$ which is $r \times 1$ and $\mathbf{y}_{it}^{(2)}$ which is $(k-r) \times 1$, so $\mathbf{y}_{it} = [\mathbf{y}_{it}^{(1)\prime}, \mathbf{y}_{it}^{(2)\prime}]'$, the OLS regression can be written as:[27]

$$\widehat{\mathbf{z}}_{it}^{+} = \mathbf{B}\mathbf{y}_{i,t-1}^{(2)} + \widehat{\mathbf{v}}_{it} \tag{2.12}$$

where now $\widehat{\mathbf{z}}_{it}^{+} = (\widehat{\boldsymbol{\alpha}}_i'\widehat{\boldsymbol{\Sigma}}_i^{-1}\widehat{\boldsymbol{\alpha}}_i)^{-1}\widehat{\boldsymbol{\alpha}}_i'\widehat{\boldsymbol{\Sigma}}_i^{-1}\Delta\mathbf{y}_{it} - \mathbf{y}_{i,t-1}^{(1)}$. Breitung then proves that this estimator has an asymptotic normal distribution. Besides extending the simple model above to more general VAR(p) models with deterministic terms, such as constant, trend and dummy variables, the Breitung (2005) procedure is more robust to contemporaneous correlation across individuals. It uses the panel-corrected standard errors suggested by Breitung and Das (2005), which correct for any effect of cross-section dependence on standard errors, whilst the estimator itself remains consistent (as T and $N \to \infty$) in the presence of such dependence. In Monte-Carlo studies, both Breitung (2005) and Wagner and Hlouskova (2007) find that this two-step estimator performs better than alternatives such as FMOLS, and the cross-sectional average of conventional cointegrated time-series estimates, when the T dimension of the panel is small (at least $T = 15$ in the former and $T = 25$ in the latter).

Thus it would seem that the Breitung two-step estimator offers an attractive solution to the question of estimating long-run cointegrating vectors. First, it makes it possible to exploit the panel dimension for the long-run coefficients only, whilst allowing heterogeneity in other coefficients such as the error-correction feedback, variance and deterministic terms, which is in the same spirit as the Pooled Mean Group approach taken by Pesaran *et al.* (1999). Secondly, it allows for contemporaneous correlation across the individuals in the panel, which is an essential feature given the nature of the dataset. Finally, the method has been shown to work well with the time and individual dimensions of this panel. Consequently, this estimator is used in the empirical work below.

[27] cf. *op. cit.* equation (6), p. 156.

Panel unit root tests

Having established how to derive estimates of long-run relationships in a panel setting, this section finishes by considering how to test for the nonstationarity that underpins the theory of cointegration. As the earlier discussion of time-series techniques for unit root testing noted, a common strategy is to use an Augmented Dickey-Fuller (ADF) test to establish the order of integration of a series. Extending this to panel models, Breitung and Pesaran (2008) distinguish between first and second generation unit root tests, on the basis that only the latter set have started to allow for cross-section dependence, characterising a similar split in approach that was observed above for cointegration methods. They set out a basic unit root test using a simple autoregressive model for y_{it}:[28]

$$\begin{aligned} y_{it} =& (1-\alpha_i)\mu_i + \alpha_i y_{i,t-1} + \epsilon_{it} \\ & i = 1,\ldots,N \quad t = 0,\ldots,T \end{aligned} \tag{2.13}$$

with $\epsilon_{it} \sim \mathsf{IID}\left(0, \sigma_i^2\right)$, $\mathsf{E}\left(\epsilon_{it}^4\right) < \infty$ and initial values y_{i0} given. By defining $\phi_i = (\alpha_i - 1)$ and $\widetilde{y}_{it} = y_{it} - \mu_i$, the resulting regression

$$\Delta\widetilde{y}_{it} = \phi_i \widetilde{y}_{i,t-1} + \epsilon_{it} \tag{2.14}$$

now has a familiar Dickey-Fuller form, and Breitung and Pesaran highlight the null (H_0) and alternative (H_{1a}, H_{1b}) hypotheses that are of interest:

$$\mathsf{H}_0 \ : \quad \phi_1 = \cdots = \phi_N = 0$$

against

$$\begin{aligned} \mathsf{H}_{1a} \ &: \quad \phi_1 = \cdots = \phi_N \equiv \phi \quad \text{and } \phi < 0 \\ \mathsf{H}_{1b} \ &: \quad \phi_1 < 0, \ldots, \phi_{N_0} < 0, \quad N_0 \leq N \end{aligned}$$

In terms of alternative hypotheses permitted under the various tests available, it is evident that H_{1a} imposes homogeneity on the alternatives, whilst

[28] See *op. cit.* pp. 4-5.

H_{1b} allows for greater heterogeneity in a set of N_0 of the N individuals.

The extent of the present interest in unit root test methods is constrained, by the likely presence of contemporaneous correlation in the dataset, to the second-generation test literature, and so the reader interested in earlier tests is directed to Breitung and Pesaran (2008) and Hlouskova and Wagner (2006) for comprehensive reviews. The literature on second-generation tests is still growing (see the introductory discussion of the former study), but three distinct approaches are emerging, and they build on a general representation of (2.13) or (2.14), in which:[29]

$$\Delta \mathbf{y}_t = \mathbf{a} + \boldsymbol{\phi} \mathbf{y}_{t-1} + \boldsymbol{\epsilon}_t \tag{2.15}$$

where the individuals from (2.13) are stacked vertically, and an arbitrary element a_i of $\mathbf{a}$ is equal to $-\phi\mu_i$. In this specification, the covariance matrix of the errors $\boldsymbol{\epsilon}_t$ is non-diagonal, so $\boldsymbol{\Omega} = \mathsf{E}\left(\boldsymbol{\epsilon}_t \boldsymbol{\epsilon}_t'\right)$ for all t, thereby allowing for cross-section dependence.

The first strategy is to treat (2.15) as a set of seemingly unrelated regressions (SUR) and then use a generalised least squares (GLS) estimator based on an estimate of the covariance matrix $\widehat{\boldsymbol{\Omega}}$. However, this requires that $T > N$, or else $\boldsymbol{\Omega}$ is singular, as Breitung and Das (2005) point out, which renders the method invalid for the full panel of 143 individuals across 15 time periods.

However, an alternative method developed by Jönsson (2005) and Breitung and Das (2005) tests H_0 against H_{1a} using OLS, with panel-corrected standard errors, which are essentially standard errors robust to cross-section correlation over the individuals (see Jönsson (2005, Section V)). With this approach, the test on ϕ has a robust t statistic, and Breitung and Das (2005) show that it has a standard normal limiting distribution under the null hypothesis H_0.

The third theme in second-generation tests builds on the presence of one or more common factors in the errors ϵ_{it}. In the simple case of one factor,

[29] This presentation draws on Breitung and Pesaran (2008, p. 20).

the error in (2.14) might be expressed as

$$\epsilon_{it} = \lambda_i f_t + v_{it}$$

where f_t is an *unobserved* effect common to *all* individuals, but varying across time, and v_{it} is an individual idiosyncratic error. In this situation, Pesaran (2007) argues that f_t can be proxied by the cross-section mean of y_{it}, and its lagged values. In the case of serially uncorrelated ϵ_{it}, H_0 is based on a test of b_i in the *cross-section* augmented Dickey-Fuller regression (CADF):[30]

$$\Delta y_{it} = a_i + b_i y_{i,t-1} + c_i \overline{y}_{t-1} + d_i \Delta \overline{y}_t + e_{it},$$

where e_{it} is white noise, and $\bar{\cdot}$ denotes a cross-section average.[31] Pesaran shows that this test, applied to the average of the t-ratios of the estimates $\left\{\widehat{b}_i\right\}_{i=1}^{N}$ allows for heterogeneity under the alternative hypothesis H_{1b}. Thus if it is the case that the contemporaneous dependence comes into the panel through a common factor, the CADF test has an advantage over the Breitung and Das method, since the panel-consistent standard errors break down, as Breitung and Pesaran (2008) observe. However, since it is not known *ex ante* how any cross-section dependence arises, it would seem pragmatic to adopt an eclectic approach and use both methods, since they are compatible with the T dimension of our panel.

In summary, this section has identified the elements of a long-run modeling strategy for the panel dataset. Bearing in mind that the individuals in the panel are likely to be contemporaneously dependent, the discussion has shown a role for unit root tests and cointegration estimators that are robust to this, and techniques that are suitable to the N and T dimensions available. Thus the following section considers how to complete the modeling strategy by exploring methods of estimating the short-run error-correction representation of the trade equations.

[30] See Pesaran (2007), equation (6).

[31] For example, $\overline{y}_{t-1} = \frac{1}{N}\sum_{i=1}^{N} y_{i,t-1}$.

2.2.2.2 Short-run estimation in dynamic panels

Why might it be interesting to explore a panel analogue of the short-run representation in (2.3)? On a general level, the rationale for doing so mirrors that of the time-series case, in which there is a distinction between long-run and short-run trade elasticities. There are good reasons why there might be differences between the two: whilst the long-run relationship is determined by more 'structural' factors such as the development of technology and tastes, the short-run relationship captures dynamic adjustment towards the long run, in the face of shocks to the explanatory variables, or to the error term. Thus there is no obvious reason to expect the elasticities across each time frame to be the same, and there is a further question of how much homogeneity exists in the dynamics across countries or sectors in the short run, versus the long run. As discussed above, the panel dimension of the dataset can be used to impose or test various homogeneity restrictions on parameters, in order to overcome some of the problems associated with a small-T sample. Exploiting similarities in technology within particular industries, for instance, could suggest imposing assumption H_{I} and pooling data across countries, allowing for heterogeneity across industries. However, whilst this might be justified when examining long-run responses, where (as Pesaran *et al.* (1999) point out) there is enough time for, say, common technology to be implemented across countries, it is possible that there is more idiosyncrasy in short-run responses. More concretely, H_{I} could determine the long-run elasticities, whilst $\mathsf{H}_{\mathrm{HET}}$ governs the short-run dynamics.

How might this be implemented empirically? The answer can be found by considering the bearing of the estimated long-run parameters from the previous section, on approaches to short-run estimation. In a single-equation environment, one option (abstracting from concerns about cross-section dependence) is to embed the long-run relationship in a reparameterised ARDL model, as done by Pesaran *et al.* (1999) in (2.9), but in view of the earlier discussion of methods to estimate long-run cointegrating vectors, an alternative avenue is one in which the long-run estimates are obtained first, and

then inserted into a short-run representation. This follows the spirit of a panel implementation of an Engle and Granger (1987) two-step procedure, based on the error-correction model shown in (2.3), which gives an $\mathsf{I}(0)$ representation of the $\mathsf{I}(1)$ cointegrated variables. In a panel setting, the process would amount to deriving estimators of the long-run cointegrating vectors (e.g. $\widehat{\boldsymbol{\beta}}$ from equation (2.11)), and then estimating a regression with $\mathsf{I}(0)$ variables.[32]

A single-equation model similar to (2.3), can demonstrate how each trade equation can be estimated in practice. For notational simplicity, y_{ist} is defined as as either imports (m_{ist}) or exports (x_{ist}), and the equation conditions on a vector $\mathbf{x}_{ist}$ where $\mathbf{x}_{ist} = (\mathsf{rpm}_{ist}, \mathsf{yd}_{ist})'$ or $\mathbf{x}_{ist} = (\mathsf{rpx}_{ist}, \mathsf{yf}_{ist})'$ as appropriate, with the vector $\mathbf{z}_{ist} = (y_{ist}, \mathbf{x}'_{ist})'$. Then, the parameters in the regression below should strictly have additional m or x subscripts to signify whether they relate to imports or exports, but in the interest of clarity these are omitted; since the estimation exercise for imports is identical to that of exports, there is no loss of understanding from doing this. The general $\mathsf{I}(0)$ specification can therefore be written as:

$$\Delta y_{ist} = \mu_{is} + \sum_{j=1}^{p} \delta_{isj} \Delta y_{is,t-j} + \sum_{j=0}^{q} \boldsymbol{\lambda}'_{isj} \Delta \mathbf{x}_{is,t-j} + \alpha_{is} \widehat{\boldsymbol{\beta}}' \mathbf{z}_{is,t-1} + v_{ist} \tag{2.16}$$

$$i = 1, \ldots, N \quad s = 1, \ldots, S \qquad t = \max(p, q) + 1, \ldots, T$$

where $\mathsf{E}(v_{ist}) = 0$, $\mathsf{E}(v_{ist}^2) = \sigma_{is}^2 < \infty$, and α_{is}, μ_{is}, $\boldsymbol{\lambda}_{isj}$ and δ_{isj} can vary across individuals. In this case, $\widehat{\boldsymbol{\beta}}' \mathbf{z}_{is,t-1}$ is the error-correction component (corresponding to $\widehat{ECM}$ in (2.3)), so that α_{is} captures the individual-specific disequilibrium feedback. Different assumptions about homogeneity in short-run parameters amount, in this setting, to restrictions on μ_{is}, δ_{isj}, $\boldsymbol{\lambda}_{isj}$, and α_{is}, and are quite distinct from assumptions governing the estimate $\widehat{\boldsymbol{\beta}}$ from

[32] The particular way of obtaining estimates of $\boldsymbol{\beta}$ determines how close this approach lies to the Engle and Granger procedure. Thus, for instance, a pure panel Engle-Granger (PEG) method would pool all data and estimate the long-run parameters from an equation in levels. The favoured approach here, though, uses the Breitung method to obtain $\widehat{\boldsymbol{\beta}}$, although the empirical discussion below does compare results to the PEG case.

the earlier I(1) analysis.

Besides this question of homogeneity, two main econometric issues are relevant to the choice of estimator. The first is the presence of individual-specific unobserved fixed effects in (2.16) in the case where homogeneity is imposed on the μ_{is} terms, so $\mu_{is} = \mu$ for all i, s. It follows from this restriction, that the error term v_{ist} can be decomposed into error components:

$$v_{ist} = \eta_{is} + e_{ist}$$

where e_{ist} is an IID error with zero mean and constant variance. If there is correlation between the individual effect η_{is} and either $\Delta \mathbf{x}_{ist}$ (and its lags) or $\mathbf{z}_{is,t-1}$, then an OLS estimator could be inconsistent. Secondly, looking at the error v_{ist} from a different perspective can reveal a different set of problems, if there is cross-sectional dependence in the panel. As Phillips and Sul (2003) comment, in this case "the pooled ordinary least squares estimator provides little gain in precision compared with single equation OLS when cross sectional dependence occurs but is ignored in the panel regression."

All of these problems – heterogeneity, cross-section dependence and fixed effects – have been addressed in the literature, so the next step is to review how they are dealt with in the set of estimators used in the empirical modeling exercise in Section 2.3.[33] This discussion is structured according to whether there is an assumption of complete heterogeneity in the short-run coefficients, against an alternative of complete (or partial) homogeneity.

In a heterogeneous world, one obvious method, given the earlier discussion of long-run approaches, is the Mean Group (MG) OLS estimator that Pesaran and Smith (1995) demonstrate to be consistent in the presence of genuine parameter heterogeneity. In practice, the basic mean group approach entails estimating (2.16) separately across all NS individuals, and then taking the cross-sectional average of each coefficient, although in view

[33]Phillips and Sul (2003) provide an elegant approach that deals jointly with both heterogeneity and cross-section dependence, but it is only valid with larger T-dimension sample sizes.

of the short time span of data, the median of estimates, which is less vulnerable to extreme observations, might also be of interest. As Pesaran and Smith show, this approach gives a meaningful value for the average effect of each independent variable on Δy_{is}. Further, differences across groups – such as countries or sectors – can be explored by calculating separate means across N and S individuals. This would correspond to pursuing the H_{I} and H_{C} hypotheses respectively, and could be useful in giving a more disaggregated impression of average effects, whilst still allowing for heterogeneity on an N-S individual level.

Although the straightforward MG estimator for (2.16) will work well if the error v_{ist} is a simple IID process as specified, it will encounter problems if there is cross-section dependence. One approach to this has been to use factor models to account for unobserved factors that vary over time, but are common to all individuals in the panel, in the same spirit as Pesaran (2007)'s CADF method for unit roots. Thus Pesaran (2006) proposes a Common Correlated Effects (CCE) estimator, which works by including the cross-section average of *all* variables, including the dependent variable, as extra regressors in (2.16), to act as proxies for the unobserved factors. This elegant solution works for multiple unobserved factors, and has the attractive property of being straightforward to implement. The resulting CCE-mean group (CCEMG) estimator is still given by the cross-section average of the estimated parameters, as before, but the regression is simply augmented with cross-section means of Δy_{ist}, $\Delta \mathbf{x}_{ist}$ and $\widehat{\boldsymbol{\beta}}' \mathbf{z}_{is,t-1}$.

Under the assumption of complete homogeneity, the simplest estimator is the pooled OLS (POLS) approach which, in contrast to the MG estimator, stacks all NS country-sector individuals vertically, and then estimates a common set of short-run parameters via OLS. This basic approach, though, is vulnerable to both unobserved individual-specific fixed effects, and cross-section dependence across individual errors. In response to the latter, Pesaran (2006) develops a pooled CCE (or CCEP), which estimates a pooled OLS regression including the same cross-section means as in the CCEMG case, as further regressors. This offers an attractive solution to the problem

of cross-section dependence in situations where alternatives might not exist – for example, if $N > T$, where a feasible GLS estimator is not available, since the estimated covariance matrix is not invertible.

Unfortunately, both the CCEP and CCEMG estimators are not robust to individual-specific unobserved heterogeneity in the form of a fixed effect, η_{is}, in the composite error v_{ist}. One solution is the within group estimator (WG), in which variables are measured in terms of their deviations from time averages. Letting $\widetilde{\cdot}$ denote this transformation, the transformed regression is:[34]

$$\Delta\widetilde{y}_{ist} = \sum_{j=1}^{p} \delta \Delta\widetilde{y}_{is,t-j} + \sum_{j=0}^{q} \boldsymbol{\lambda} \Delta\widetilde{\mathbf{x}}_{is,t-j} + \alpha(\widehat{\boldsymbol{\beta}}' \widetilde{\mathbf{z}}_{is,t-1}) + \widetilde{v}_{ist} \tag{2.17}$$

This case allows for an individual (i.e. country-sector) specific fixed effect η_{is}, which is eliminated by expressing variables in deviations from time means. However, the price of this is that some degree of homogeneity is imposed on δ, $\boldsymbol{\lambda}$ and α. In (2.17) as it stands, the parameters are assumed to be fixed across all NS individuals. However, it would be simple to allow for heterogeneity across countries or sectors, by running $N = 11$ or $S = 13$ separate pooled WG regressions, respectively.

One issue in a within group model, is the possibility that the presence of lagged dependent variables (either through lagged Δy_{is}s or the inclusion of $y_{is,t-1}$ in $\widehat{\boldsymbol{\beta}}' \mathbf{z}_{is,t-1}$) could introduce bias for small/fixed T, as explored by Nickell (1981). If this were the case then the α_{is} estimates would be biased downwards, but given $T = 13$ (using one lagged Δy_{is} or $\Delta \mathbf{x}_{is}$) it seems unlikely that such bias will be significant here. More likely as a concern is the presence of cross-section dependence, which can be addressed via the Pesaran (2006) CCE estimator, augmenting the WG regression (2.17) with cross-section averages, this time adjusted for individual-specific effects.

[34] Thus for example, when $p = 1$ and $q \leq 1$, $\Delta\widetilde{y}_{ist} = \Delta y_{ist} - \frac{1}{T-1} \sum_{j=2}^{T} \Delta y_{isj}$.

2.3 Empirical Modeling

The empirical results presented below suggest that there is a long-run relationship between the variables in each trade flow. Cointegration tests indicate that this relationship is strongest across industry groups, which supports the H_{I} hypothesis that long-run elasticities are common to industries across countries. In contrast, the evidence for homogeneity *within* countries is weaker, whilst the most restrictive case of complete homogeneity across all country-industry pairs is rejected.

Using industry-by-industry estimates of the long-run relationship, the short-run analysis also finds that group-by-group regressions, and the more flexible Mean Group estimator, deliver more robust estimates than the fully pooled regressions.

2.3.1 Testing for integration and cointegration

The earlier discussion of panel unit root tests highlighted the need for robustness to potential cross-section dependence, so this section uses both the Breitung and Das (2005) test of a homogeneous unit root against an alternative homogeneous stable root, and the approach developed by Pesaran (2007) which tests the same null hypothesis against a heterogeneous alternative. For reference, these results are compared to those from an earlier vintage of unit root tests (which were not robust to cross-section dependence), via the method of Im, Pesaran and Shin (2003).

Table 2.11 in Appendix 2.B.1 presents persuasive evidence in favour of all the variables having a homogeneous unit root. With one exception (foreign income in the CADF(1) test), all the test statistics for the robust CADF and BD tests suggest that the data are $\mathsf{I}(1)$, and the importance of accounting for cross-section dependence is underlined by the contrast with column 3, which reports the IPS test statistics. With this method, a unit root is unanimously rejected for half the series considered.

An obvious next step is to test for the presence of *co*integration between the variables, and the results from Breitung (2005)'s rank test on $\mathbf{\Pi}_i = \boldsymbol{\alpha}_i \boldsymbol{\beta}'$

in (2.11) are presented in Table 2.12 in Appendix 2.B.1. The rank test is based on the standardised average of individual-specific LM test statistics, with a null hypothesis of up to r cointegrating relationships, against an alternative of no cointegration (i.e. $r = 0$). Since this is a 'sequential' test, if a null hypothesis (of, say, $r = 1$) is accepted, then that value of r represents the true number of cointegrating relationships, so no further tests are necessary.

Table 2.12 is divided into three panels, each of which reflects a different pooling assumption: Panel (a) refers to the common elasticities case ($\mathsf{H}_{\text{NOHET}}$), (b) refers to H_{C}, and (c) refers to H_{I}. The first of these suggests that the hypothesis of one cointegrating vector for each trade flow across the whole sample is rejected. By imposing the same relationship on all country-industry pairs, the heterogeneity that seems to exist across groups is forced into the unobserved error term in the pooled cointegrating regression. The evidence in panel (b) is less clear-cut, as whilst there is support for one cointegrating vector in a number of countries, there is also evidence of no cointegration in 6 out of the 22 tests. Against this picture, panel (c) is nearly unanimous in indicating that there is one long-run relationship in each of the industries tested, with the null hypothesis rejected only twice, and in one of these, only marginally.

Taken together, these results point to the existence of one cointegrating vector in each sector, which is consistent with the theoretical justification discussed earlier, relating to similarities in production functions and technology across industries.

2.3.2 Long-run elasticities

The main empirical results of this chapter are the long-run elasticities presented in Tables 2.3 to 2.5, which correspond to the three homogeneity restrictions $\mathsf{H}_{\text{NOHET}}$, H_{C} and H_{I}, respectively. The first theme that emerges from these results is that the aggregate approach embodied in the common elasticities in Table 2.3 disguises considerable heterogeneity in the group-by-group analysis in the remaining two tables. As the latter report, the stan-

Table 2.3: LONG-RUN TRADE ELASTICITIES: HOMOGENEOUS CASE

Sample	Exports (x)		Imports (m)	
	Price (rpx)	Income (yf)	Price (rpm)	Income (yd)
Common elasticities	−0.927	1.509	−0.861	1.715
	(0.027)	(0.037)	(0.028)	(0.032)

NOTES: Estimation period is 1988-2002 using annual data. Estimates obtained via the Breitung two-step estimator in (2.12), with long-run coefficients normalised on exports (x) or imports (m) as appropriate. Figures in (·) denote standard errors. In the case of complete pooling, data have been pooled across all $NS = 143$ country-sector individual units, for $T = 15$ time periods. *GAUSS code:* `pan2step.src` (kindly provided by Joerg Breitung); *Ox code:* `DataOverview.ox`.

dard deviations of all trade elasticities are substantial, although the mean and median estimates are reasonably close to the common pooled figures.

Thinking about the wider purpose of this chapter, these estimates can be compared to the selected estimates in Tables 2.1 and 2.2, to see what additional insight is gained from the panel approach.

On a country-by-country basis, the panel-based estimates show greater consistency, in the sense that the estimates are all significant (with the sole exception of Norway's export price elasticity), of plausible magnitude, and have signs consistent with economic theory; in contrast, many of the earlier estimates are insignificant (especially for price elasticities) and several have incorrect signs. The comparison with more recent VECM-based estimates from Hooper *et al.* (2000) is more favourable, with similar patterns in price elasticities, although the panel-based income coefficients seem to be higher in many cases.

However, a more fundamental comparison to make is between the country-by-country and industry-by-industry results, as the latter point to significant variation in elasticities within the sectors in a country. Irrespective of whether the former are consistent with existing estimates in the literature, it is inadvisable to ignore potential within-country variation.

Looking at Table 2.5, it is worth picking out some patterns in the evidence. Starting with outlying estimates, it is interesting that both Chemical

Table 2.4: LONG-RUN TRADE ELASTICITIES: COUNTRY HETEROGENEITY

Sample	Exports (x)		Imports (m)	
	Price (rpx)	Income (yf)	Price (rpm)	Income (yd)
Belgium	−0.693	1.850	−0.624	2.226
	(0.060)	(0.087)	(0.133)	(0.114)
Canada	−0.948	2.696	−0.979	2.058
	(0.091)	(0.096)	(0.048)	(0.070)
Denmark	−0.564	1.166	−0.742	1.902
	(0.067)	(0.085)	(0.081)	(0.075)
Finland	−0.906	2.120	−0.853	2.010
	(0.149)	(0.313)	(0.065)	(0.086)
Germany	−0.854	1.063	−1.300	1.979
	(0.059)	(0.072)	(0.121)	(0.126)
Italy	−0.610	2.003	−0.854	3.220
	(0.093)	(0.147)	(0.053)	(0.103)
Japan	−0.952	0.530	−1.152	3.667
	(0.066)	(0.082)	(0.076)	(0.232)
Netherlands	−0.577	1.188	−2.126	1.150
	(0.131)	(0.161)	(0.121)	(0.076)
Norway	−0.252	0.783	−1.068	0.533
	(0.154)	(0.112)	(0.113)	(0.075)
UK	−1.038	1.690	−1.178	0.929
	(0.088)	(0.092)	(0.068)	(0.096)
US	−1.192	2.209	−0.856	2.396
	(0.055)	(0.096)	(0.063)	(0.056)
Unweighted Average	−0.781	1.573	−1.066	2.006
(standard deviation)	(0.256)	(0.642)	(0.386)	(0.880)
Median	−0.854	1.690	−0.979	2.010

NOTES: Estimation period is 1988-2002 using annual data. Estimates obtained via the Breitung two-step estimator in (2.12), with long-run coefficients normalised on exports (x) or imports (m) as appropriate. Figures in (·) denote standard errors. For the country-specific estimates, separate regressions have been run for each country, using data pooled across all $S = 13$ industries, across $T = 15$ time periods. *GAUSS code:* `pan2step.src` (kindly provided by Joerg Breitung); *Ox code:* `DataOverview.ox`.

Table 2.5: LONG-RUN TRADE ELASTICITIES: INDUSTRY HETEROGENEITY

Sample	Exports (x)		Imports (m)	
	Price (rpx)	Income (yf)	Price (rpm)	Income (yd)
Food, beverages and tobacco	−0.739	1.430	−0.532	1.240
	(0.078)	(0.111)	(0.059)	(0.060)
Textiles, leather and footwear	−0.668	1.292	−0.475	1.313
	(0.152)	(0.178)	(0.109)	(0.097)
Wood and wood products	−0.869	1.015	−0.448	1.444
	(0.107)	(0.173)	(0.078)	(0.102)
Paper, printing and publishing	−0.737	0.858	−0.168	0.985
	(0.078)	(0.082)	(0.063)	(0.062)
Chemical products	−0.751	2.081	−0.066	1.941
	(0.062)	(0.077)	(0.099)	(0.088)
Rubber and plastics	−0.978	1.598	−0.425	1.685
	(0.106)	(0.131)	(0.101)	(0.086)
Non-metallic minerals	−0.849	1.011	−0.500	1.293
	(0.097)	(0.108)	(0.105)	(0.096)
Basic metal products	−0.706	0.811	−0.048	1.181
	(0.066)	(0.082)	(0.090)	(0.072)
Machinery and equipment	−0.842	2.446	−0.771	2.856
	(0.096)	(0.141)	(0.144)	(0.188)
Other machinery and equipment	−0.877	1.487	−0.503	1.311
	(0.088)	(0.092)	(0.098)	(0.101)
Electrical and optical equipment	−0.941	2.911	−1.234	2.660
	(0.103)	(0.222)	(0.112)	(0.261)
Transport	−0.559	1.557	−0.591	1.805
	(0.073)	(0.079)	(0.114)	(0.115)
Other Manufacturing	−0.723	1.861	−0.521	1.823
	(0.079)	(0.111)	(0.076)	(0.083)
Unweighted Average	−0.788	1.566	−0.483	1.657
(standard deviation)	(0.113)	(0.602)	(0.295)	(0.544)
Median	−0.751	1.487	−0.500	1.444

NOTES: Estimation period is 1988-2002 using annual data. Estimates obtained via the Breitung two-step estimator in (2.12), with long-run coefficients normalised on exports (x) or imports (m) as appropriate. Figures in (·) denote standard errors. For the industry-specific estimates, separate regressions have been run for each sector, using data pooled across all $N = 11$ countries, across $T = 15$ time periods. *GAUSS code:* `pan2step.src` (kindly provided by Joerg Breitung); *Ox code:* `DataOverview.ox`.

products and Machinery and equipment (and the sub-group within this, of Electrical and optical equipment) have by far the highest income elasticities of all industries. These sectors fall under the broad definition from the OECD of 'hi-tech' industries, so the difference between these estimates and those of, say, Basic metal products (where the scope for technological improvements in the latter may be lower) could be related to inputs such as research and development activity. The distinction is less clear-cut with regard to price elasticities, though, since Chemicals imports have one of the lowest estimates, and Electrical and optical equipment the highest.

Whilst the scope of this study is limited to establishing whether there *are* differences in elasticities across industries, and not *why* such differences might exist, it is clear that there is an exciting research agenda in the making. Some studies (such as Driver and Wren-Lewis (2005)) have started to look at the impact of product quality and its role in country-wide trade performance, but a wider industry-level study could explore in more detail the factors behind the patterns in Table 2.5.

The theoretical rationale for revisiting the topic of trade elasticities with industry-level panel data was that aggregate and pooled studies might disguise the existence of substantial variation within countries, whilst also affecting the statistical validity of estimated relationships. As these results show, the empirical picture supports this theoretical position. Tests using a completely pooled sample reject the hypothesis of cointegration, and whilst there is more success on a country-specific level, there are nonetheless problems. On an industry-by-industry basis, though, the evidence for cointegration is overwhelming, and the long-run elasticities are consistent with economic theory. Furthermore, they raise several questions to explore in future research.

2.3.3 Short-run elasticities

How do the long-run estimates enter into short-run estimation? The foregoing discussion provided a rationale for allowing for heterogeneity across industries in the long run, so the equilibrium-correction term in all short-run

models is given by $\widehat{\boldsymbol{\beta}}_s'\mathbf{z}_t$, which uses the industry-specific long-run elasticities.[35] Looking at the estimates of short-run elasticities in Tables 2.6 to 2.9, it is worth starting with three points about general trends.

First, the short-run price elasticities are broadly significant, and correctly signed, and suggest that much of the short-run effect of price on trade comes through the *contemporaneous* relative price, rather than lagged prices.

Secondly, the impact of income on trade is more complex, as many of the estimated coefficients are insignificant, and whilst contemporaneous income growth typically has a positive coefficient, lagged income enters negatively. Within this picture, though, it is possible to draw out two observations. Where the sum of current and lagged coefficients on imports and exports is significant, it is also positive, which supports economic theory. Further, in such cases, the import income elasticity sum is greater than that of exports.

Finally, the feedback coefficient on the equilibrium-correction term is significant in all but one case, and has an appropriate (negative) sign in all. The picture of heterogeneity in long-run elasticities is echoed by the short-run results, which suggest that the estimates from group-wise, and completely heterogeneous regressions, are more robust than those from the completely pooled scenario. There is evidence of serial correlation in the residuals from completely pooled regressions, which is consistent with the effect of ignoring, incorrectly, genuine cross-section heterogeneity in parameters, as Pesaran and Smith (1995) predicted.[36] In contrast, the alternative approaches allowing for more heterogeneity do not suffer from the same mis-specification problems.

Besides this general issue, there are specific comments to make about the results from the four separate homogeneity scenarios reported in the Tables.

Addressing the complete pooling case, the main point to note is the effect

[35]In the tables, the equilibrium-correction term is written as $\widehat{\mathsf{ECM}}$. For example, in the case of Chemical product exports (using the abbreviation s to denote the sector) this yields:

$$\widehat{\mathsf{ECM}}_{s,t} = \widehat{\boldsymbol{\beta}}_{s,\mathsf{x}}'\mathbf{z}_{\mathsf{x},t} = \mathsf{x}_{s,t} + 0.751\mathsf{rpx}_{s,t} - 2.081\mathsf{yf}_{s,t}.$$

[36]This is shown in the results of LM tests for first- and second-order serial correlation in the residuals from each regression, reported in the lower panel of Tables 2.6 to 2.9.

Table 2.6: SHORT-RUN ELASTICITIES ACROSS THE COMPLETE SAMPLE WITH COMMON CORRELATED EFFECTS: IMPORTS

Dependent variable is $\Delta \mathsf{m}_t$	(1) POLS	(2) FE	(3) 2SLS-FE	(4) MGOLS		(5) MG2SLS	
				Mean	Median	Mean	Median
$\Delta \mathsf{m}_{t-1}$	0.075	0.063	0.070	−0.081	−0.120	−0.362	−0.080
	(0.027)	(0.027)	(0.023)	(0.047)	(−)	(0.377)	(−)
$\Delta \mathsf{rpm}_t$	−0.449	−0.424	−0.605	−0.285	−0.410	−0.344	−0.560
	(0.029)	(0.029)	(0.115)	(0.095)	(−)	(0.842)	(−)
$\Delta \mathsf{rpm}_{t-1}$	−0.096	−0.072	−0.049	−0.007	−0.063	0.019	−0.024
	(0.030)	(0.031)	(0.031)	(0.079)	(−)	(0.348)	(−)
$\Delta \mathsf{yd}_t$	2.272	2.428	2.412	2.598	1.894	2.710	2.215
	(0.162)	(0.155)	(0.131)	(0.388)	(−)	(2.213)	(−)
$\Delta \mathsf{yd}_{t-1}$	−1.159	−0.980	−1.113	−0.514	−0.416	1.196	−0.173
	(0.155)	(0.152)	(0.156)	(0.323)	(−)	(2.556)	(−)
$\widehat{\mathsf{ECM}}_{t-1}$	−0.002	−0.165	−0.160	−0.358	−0.254	−0.360	−0.287
	(0.000)	(0.019)	(0.015)	(0.053)	(−)	(0.303)	(−)
Pooled data	Yes	Yes	Yes	No	No	No	No
Fixed effects	No	Yes	Yes	N/A	N/A	N/A	N/A
SC: AR(1)	0.033	0.000	0.000	0.993		0.965	
SC: AR(2)	0.043	0.000	0.000	(−)		(−)	

NOTES: Figures in (·) denote heteroscedasticity-robust standard errors for pooled estimators, and the group average standard error in the case of mean group estimates. The rows marked SC report tests for first- and second-order serial correlation in the errors. For pooled estimates (1) to (3), these are the p-value of incorrectly rejecting the null hypothesis of no serial correlation; for mean group regressions (4) and (5) these are the proportion of all $NS = 143$ tests that did not reject the null hypothesis at the 1% significance level. *Ox code:* `TradeEstimation.ox`.

of allowing for individual-specific fixed effects in estimation. Columns (1) and (2) in Tables 2.6 and 2.7 show the estimates obtained first, in an OLS regression of stacked individuals, and secondly, in the same OLS regression but including a set of individual-specific dummy variables.[37] The main impact of this difference manifests in the feedback coefficient, which jumps from −0.001 to −0.185 after allowing for fixed effects. This is consistent with

[37]These regressions also include cross-section means of all variables, following the Pesaran CCE approach. Results of regressions without the CCE estimator are available on request from the author.

Table 2.7: SHORT-RUN ELASTICITIES ACROSS THE COMPLETE SAMPLE WITH COMMON CORRELATED EFFECTS: EXPORTS

Dependent variable is Δx_t	(1) POLS	(2) FE	(3) 2SLS-FE	(4) MGOLS		(5) MG2SLS	
				Mean	Median	Mean	Median
Δx_{t-1}	0.060	0.036	0.010	−0.210	−0.124	−0.238	−0.171
	(0.035)	(0.035)	(0.025)	(0.050)	(−)	(0.141)	(−)
Δrpx_t	−0.631	−0.585	−0.302	−0.634	−0.615	−1.558	−0.640
	(0.031)	(0.031)	(0.086)	(0.082)	(−)	(1.077)	(−)
Δrpx_{t-1}	−0.150	−0.117	−0.136	−0.199	−0.115	0.589	−0.111
	(0.029)	(0.030)	(0.026)	(0.048)	(−)	(0.775)	(−)
Δyf_t	2.143	2.011	1.540	−1.008	−0.443	−0.535	−0.011
	(0.373)	(0.395)	(0.376)	(1.174)	(−)	(4.477)	(−)
Δyf_{t-1}	0.055	0.142	0.148	0.577	0.784	1.219	0.154
	(0.266)	(0.248)	(0.226)	(0.343)	(−)	(1.620)	(−)
$\widehat{ECM}_{t-1}$	−0.001	−0.185	−0.195	−0.382	−0.325	−0.519	−0.425
	(0.000)	(0.018)	(0.015)	(0.055)	(−)	(0.209)	(−)
Pooled data	Yes	Yes	Yes	No	No	No	No
Fixed effects	No	Yes	Yes	N/A	N/A	N/A	N/A
SC: AR(1)	0.020	0.000	0.001	0.958		1.000	
SC: AR(2)	0.006	0.000	0.000	(−)		(−)	

NOTES: See notes to Table 2.6. *Ox code:* `TradeEstimation.ox`.

each country-sector pair having a specific intercept, with the possibility that it lies in the cointegrating space. Since the Breitung procedure for the long-run relationships also allows for intercepts varying across individuals, the fixed effects estimator used here is a consistent short-run analogue, and so column (2) in both Tables could be seen as the preferred pooled specification, of the two considered so far.

However, the robustness of the fixed effects OLS specification is affected by potential endogeneity bias arising from correlation of the explanatory variables with the error term. Of the variables considered, the contemporaneous changes in income are unlikely to be correlated with country-sector-specific trade shocks, since the size of a particular sector relative to the aggregate scale of both foreign and domestic income is likely to be very small. Similarly, as Carlin, Glyn and van Reenen (2001) note, exchange

Table 2.8: SHORT-RUN ELASTICITIES FROM DISAGGREGATED GROUP REGRESSIONS: IMPORTS

Dependent variable is $\Delta \mathsf{m}_t$	Country groups 2SLS-FE-CCE		Industry groups 2SLS-FE-CCE	
	Mean	Median	Mean	Median
$\Delta \mathsf{m}_{t-1}$	0.068	0.055	0.020	0.005
	(0.080)	(−)	(0.039)	(−)
$\Delta \mathsf{rpm}_t$	−1.933	−0.622	−0.692	−0.349
	(2.740)	(−)	(0.293)	(−)
$\Delta \mathsf{rpm}_{t-1}$	−0.174	−0.050	−0.030	−0.163
	(0.091)	(−)	(0.071)	(−)
$\Delta \mathsf{yd}_t$	0.517	0.044	2.463	2.425
	(0.352)	(−)	(0.194)	(−)
$\Delta \mathsf{yd}_{t-1}$	−0.498	−0.180	−1.079	−0.675
	(0.359)	(−)	(0.456)	(−)
$\widehat{\mathsf{ECM}}_{t-1}$	−0.217	−0.223	−0.190	−0.225
	(0.065)	(−)	(0.030)	(−)
SC: AR(1)	1.000		0.769	
SC: AR(2)	0.909		0.846	

NOTES: The coefficients are the mean and median of separate group-by-group pooled regressions. For example, in the case of Country Groups (columns 2-4), this amounts to running 11 separate 2SLS regressions, including fixed effects and Pesaran CCE effects. The choice of estimator is directed by its robustness to the empirical problems discussed in the text. Estimates from other methods, and a complete breakdown of results by country and industry, are available by request from the author. *Ox code:* `TradeEstimation.ox`.

rates are unlikely to be affected by shocks to an individual sector. Where endogeneity could emerge, though, is through pricing behaviour affecting the price level components of the relative price terms. Thus as a check for robustness, a fixed-effects specification is reported in column (3), in which lagged values of relative price levels (rpm_{t-2} and rpx_{t-2}) are used as instruments for contemporaneous price changes (respectively, $\Delta\mathsf{rpm}_t$ and $\Delta\mathsf{rpx}_t$), and the response is estimated via two-stage least squares (2SLS). Neither the import nor export price elasticities change dramatically after doing this, suggesting that endogeneity issues are of second-order concern for the pooled case.

Table 2.9: SHORT-RUN ELASTICITIES FROM DISAGGREGATED GROUP REGRESSIONS: EXPORTS

Dependent variable is Δx_t	Country groups 2SLS-FE-CCE		Industry groups 2SLS-FE-CCE	
	Mean	Median	Mean	Median
Δx_{t-1}	−0.007	0.025	0.077	0.065
	(0.028)	(−)	(0.024)	(−)
Δrpx_t	−1.249	−1.153	−0.438	−0.312
	(0.502)	(−)	(0.139)	(−)
Δrpx_{t-1}	−0.045	−0.048	−0.047	−0.058
	(0.021)	(−)	(0.029)	(−)
Δyf_t	1.665	0.959	2.356	2.245
	(0.949)	(−)	(0.450)	(−)
Δyf_{t-1}	−0.016	−0.056	−0.978	−0.819
	(0.084)	(−)	(0.373)	(−)
$\widehat{ECM}_{t-1}$	−0.202	−0.195	−0.246	−0.219
	(0.033)	(−)	(0.031)	(−)
SC: AR(1)	0.909		0.846	
SC: AR(2)	0.909		0.769	

NOTES: See notes to Table 2.8. *Ox code:* `TradeEstimation.ox`.

The fully pooled results, whilst showing signs of mis-specification, are nonetheless useful in providing a guide to the appropriate choice of estimator for the group-wide regressions, since they suggest that where the data are pooled, a fixed-effects model should be used, and in the face of potential endogeneity, it should also be ready to correct for it.

Thus the next step in analysis is to compare the fully pooled results to the group-by-group pooled estimates in Tables 2.8 and 2.9. These are obtained by running separate regressions in which all the individuals in a group are pooled, and then the average of all group estimates is taken. For example, in the case of country-by-country results in the first block of each table, 11 sets of country-specific pooled coefficients are produced, and the mean and median are reported here.[38] In a sense, these represent pseudo-

[38]The 2SLS estimator allowing for fixed effects (via group-specific dummies), and cross-section dependence (via the CCE estimator), is used for these regressions. Full country- and industry-specific breakdowns are available for this method, and alternatives, on re-

mean group estimates where the $NS = 143$ individual OLS regressions in the conventional MG approach are replaced by 11, or 13, pooled regressions.

The most important message that emerges from these results is that allowing for heterogeneity across groups leads to a dramatic improvement in diagnostic test statistics. As the lower panels of Tables 2.8 and 2.9 show, the fraction of group-wide regressions that pass the serial correlation tests is very high, with the country-by-country results showing less sign of misspecification than the industry regressions. This lends support to the idea that the sectors in a given country behave in a similar way in the face of shocks in the short run (when shocks are more likely to hit whole countries, rather than particular industries), whilst in the long run, behaviour is common to industries *across* countries (when sectors have the ability to adopt new technology from abroad, for example).

Moving beyond the group-wide regressions, the final element of the short-run analysis reports Mean Group results, in blocks (4) and (5) of Tables 2.6 and 2.7, and Figures 2.2 and 2.3.

Although there are constraints on inference imposed by the fact that the number of time periods available ($T = 13$, once lagged changes are included) is only just higher than the number of regressors (10, including a constant and the CCE cross-section means), there are nonetheless several points to make.

First, the individual models do not suffer from the same serial correlation as their fully pooled counterparts in columns (1) to (3), which supports the hypothesis that the diagnostic failure of the latter is due to parameter heterogeneity.

An immediate question that follows from this is whether the superior diagnostic performance is accompanied by any substantial differences in coefficient estimates. In this regard, the mean and median of individual-specific feedback coefficients on the equilibrium-correction terms are significantly higher across MG estimates than any of the pooled alternatives, and as the top right-hand panel of Figures 2.2 and 2.3 show, the majority if estimates

quest from the author.

Figure 2.2: MEAN GROUP OLS SHORT-RUN IMPORT COEFFICIENT ESTIMATES: COMPLETE HETEROGENEITY

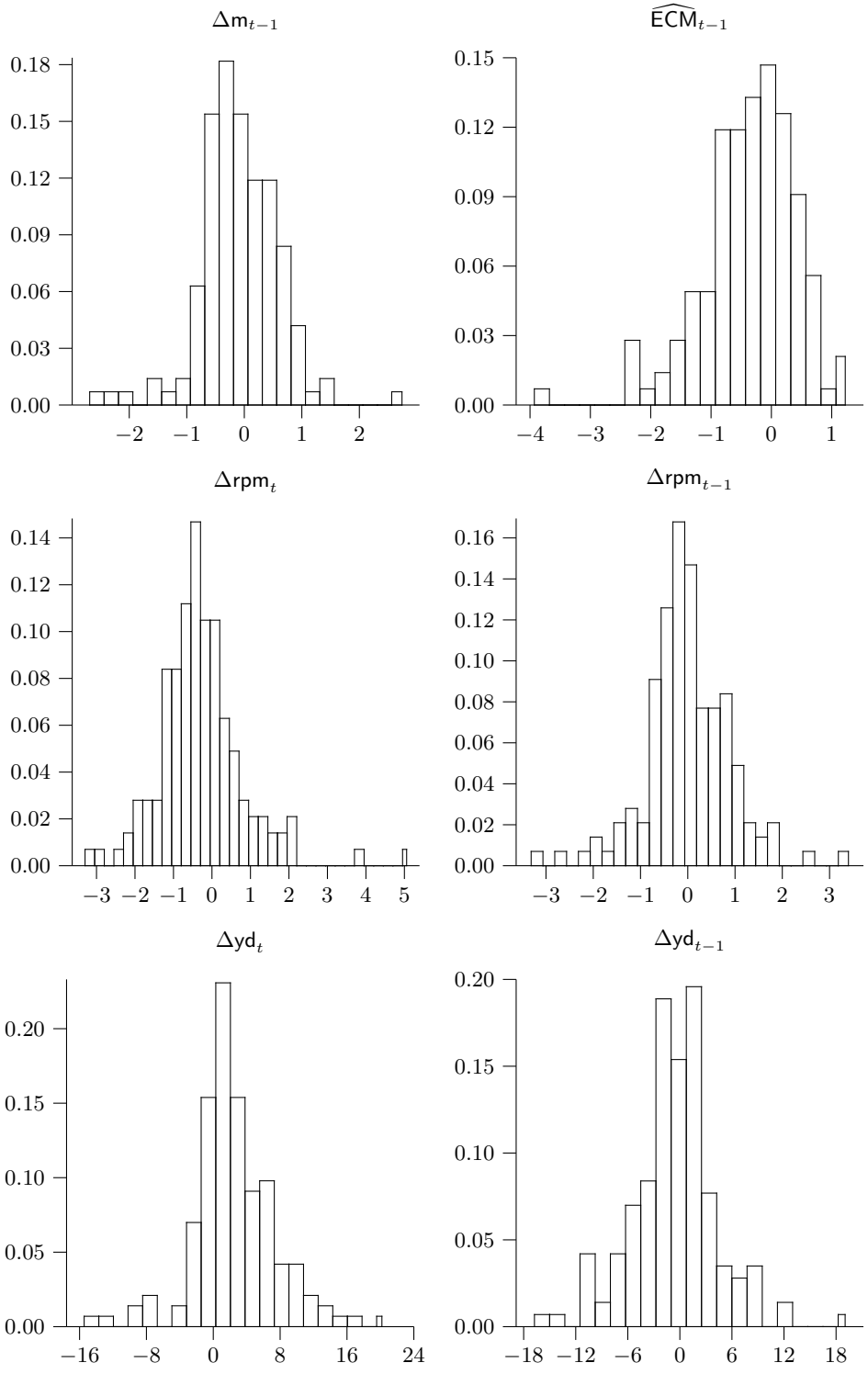
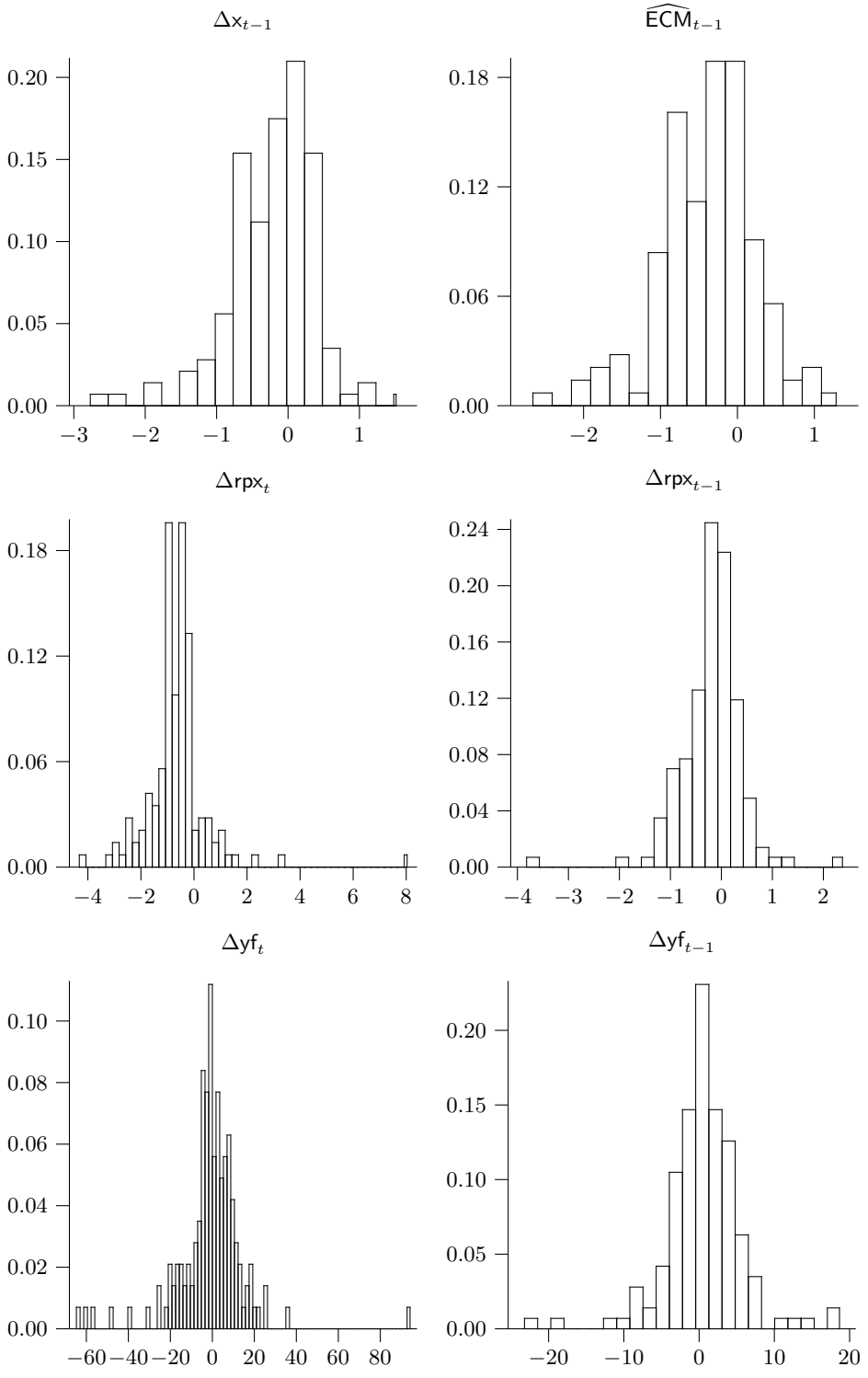


Figure 2.3: MEAN GROUP OLS SHORT-RUN EXPORT COEFFICIENT ESTIMATES: COMPLETE HETEROGENEITY



lie between 0 and −1.

Thirdly, the robustness check for endogeneity on an individual level via the 2SLS regression reported in block (5) does not point to significant differences in the relevant elasticities, although the standard errors obtained in this case are substantially higher, which affects the precision of any judgment about the impact of endogeneity.

Finally, the distribution of estimates in Figure 2.2 and 2.3 provides an idea of how plausible the MG results are, given the small dimension of the sample. In broad terms, the behaviour of the elasticity coefficients corresponds to the general trends seen in pooled estimates. Looking at the middle row of both Figures, the distributions of price elasticities are generally skewed towards negative values, although there is some evidence of positive coefficients for lagged import prices. With regard to income elasticities in the bottom row, the distribution of estimates on contemporaneous income growth is heavily skewed towards positive values, whilst those of lagged growth are more evenly distributed around zero, echoing the pooled coefficient results. Besides some extreme outliers (e.g. an export income elasticity of 90), the overall picture painted by the MG estimates seems to follow that of the group-wise pooled regressions, with the exception that the responsiveness of trade to long-run disequilibrium is higher in the heterogeneous case than in the alternative approaches.

2.4 Concluding comments

These results augur well for a wider research agenda on disaggregated trade elasticities. First, a disaggregated view was shown to be necessary to understand how imports and exports in a particular country will respond to a change in the exchange rate. To see why, consider an aggregate measure of the exchange rate, such as the effective exchange rate index (ERI), which weights a country's bilateral exchange rates according to each partner's share of total trade. An appreciation in the ERI may come about due to movements in any one of the bilateral rates, and if some industries trade more with a particular partner than others, then the same movement in the

ERI can have different effects on industries, depending on which exchange rate is moving.

Secondly, the framework established above can readily be extended to include additional factors such as innovation or research and development (as explored by Driver and Wren-Lewis (2005)), which would allow wider questions about the determinants of trade performance to be analysed.

Finally, extending the time span of the database could open an avenue to examining the long-run impact on trade performance, of short-run shocks. Thus if an exchange rate shock changes the properties of the long-run cointegrating vector (such as its mean, or the income and price elasticities), then analysis of the impact of such breaks over time would offer insight into the response of individual sectors to macroeconomic shocks (as studied, for example, by Kitson and Primost (2005)).

2.A Data Appendix

2.A.1 STAN

The Structural Analysis (STAN) Database from the OECD is compiled from member countries' annual national accounts, national surveys and censuses and other sources needed to fill any missing entries. In this chapter the latest version of STAN is used, containing data running up to 2003 (see OECD (2005, 2006*b*))[39] and data comes from two parts of the database:

- Real and Nominal Value Added, and Nominal Imports and Exports, all in national currencies, are taken from the main STAN Database;[40]
- Bilateral trade data, which measures Nominal Imports and Exports broken down by trading partner, for each country-industry unit, are taken from the STAN Bilateral Trade Database (BTD).[41]

Drawing on the approach of Carlin *et al.* (2001), the industry coverage of the empirical work in this chapter is set out in Table 2.10, which indicates

[39] Although the empirical work here uses data up to 2002, due to partial country-sector coverage of the data for 2003.

[40] OECD (2005): see http://www.oecd.org/sti/stan/.

[41] OECD (2006*b*): see http://www.oecd.org/sti/btd/.

Table 2.10: STAN INDUSTRY COVERAGE

Overall Group	ISIC	Group breakdown
Food, beverages and tobacco	*15-16*	Food and beverages*; tobacco*
Textiles, leather and footwear	*17-19*	Textiles*; Leather and footwear*
Wood and wood products	*20*	Wood and wood products
Paper, printing and publishing	*21-22*	Pulp and paper*; Printing and Publishing*
Chemical products	*23-25*	Coke, refined petroleum and nuclear fuel*; Chemical products* (non-pharmaceutical chemicals*; pharmaceuticals*); rubber and plastics
Non-metallic minerals	*26*	Non-metallic mineral products
Basic metal products	*27-28*	Basic metals* (iron and steel*; other metals*); Fabricated metal products excluding machinery and equipment*
Machinery and equipment	*29-33*	Electrical and optical equipment (office and computing*; radio, television and communication*; medical equipment*; other equipment*); other machinery and equipment
Transport	*34-35*	Motor vehicles*; Other transport* (ships and boats*; aircraft*; railroad equipment*)
Other manufacturing	*36+37*	Furniture*; other manufacturing*; recycling*

NOTES: Source: OECD (2005). The ISIC codes refer to the two-digit classification under ISIC Revision 3. An * indicates that the particular group breakdown was not used due to lack of data across all countries in the sample.

the highest level of disaggregation available in STAN, and the level actually used in this work.

2.A.2 INDUSTRY, TIME AND COUNTRY COVERAGE OF STAN

The criteria for selecting variables are as follows:

- The variables of interest are nominal imports and exports, and nominal and real value added. Using the latter pair, deflators can be constructed for both value added, and imports and exports. In order to construct trade-weighted variables where relevant, *bilateral* trade data, which decomposes trade for a particular country's sector into partner countries of origin (for imports) or of destination (for exports), are

taken from the STAN Bilateral Trade Database (BTD).[42]

- Although the main STAN database covers all areas of activity, there are practical constraints on the number of sectors that can be used. First, the BTD data, which are essential in calculating trade weights, only cover manufacturing sectors at present. This is clearly a limitation that future versions of BTD could resolve, but for now, the effect is that the empirical work in this chapter focuses on manufacturing industries. Within these, activity is disaggregated into groups according to the International Standard of Industrial Classification of Economic Activities (ISIC), revision 3. This has the advantage of providing a consistent definition of each sector, making comparison across countries more meaningful, although as Carlin *et al.* (2001) note, "... because of the way in which the data set was constructed, the extent of measurement error increases the more disaggregated the data" (p. 131). For this reason, in combination with data availability, the sectoral data used here is disaggregated to the ISIC two-digit level. In total, this leaves nine broad industry groups and a residual manufacturing sector, of which some of these are split into subgroups (still at the two-digit level), so that the S (industry) dimension of the panel comprises 13 'individuals.'[43]

- Although STAN covers 29 countries, there are two concerns about including all in the panel dataset. First, echoing Carlin *et al.* (2001), some countries, such as Mexico and Korea, were at relatively low levels of development in the early period of the sample, and so their experiences would partly reflect an adjustment process not shared with the others. Secondly, data availability affects a number of countries, leaving 11 to include in the dataset.

- The time dimension of the panel, once the countries and sectors have been selected, covers fifteen years, from 1988 to 2002. Since the data

[42]The Bilateral Trade Database is available from http://www.oecd.org/sti/btd.
[43]The full list of industries is presented in Table 2.10.

has annual frequency, it is immediately clear that conventional time-series approaches will have very low power in such a small sample. Thus the empirical work below, makes use of the panel dimension in a number of ways.

2.A.3 Variable definitions

The variables used in the empirical modeling are as follows:[44]

- Imports, denoted m_{ist} measure the real value of all imports of goods in sector s into country i, in year t. This has been deflated by the GVA deflator in each partner country, with weights to account for the share of each partner in total s imports into country i.

- Exports, denoted x_{ist} measure the real value of all exports from sector s in country i. The nominal value of exports is deflated using the GVA deflator for that particular sector's output, since an explicit export price measure is not available in the STAN database; under the assumption that the overall GVA deflator for a particular sector provides an appropriate measure of the price of exported goods for that sector, this derived index is a suitable one to use.

- Relative import prices, rpm_{ist}, measure the price of imported goods from sector s, adjusted by the exchange rate, relative to the price of *domestic output* from sector s. This definition is closest to the most accurate theoretical measure of the robust relative price series that was identified in Section 2.1. The trade-weighted exchange rate measure is constructed using bilateral exchange rates between each country and its trading partners, weighted by the share of each partner in the imports for each sector. This provides sector-specific exchange rate indicies.

- Relative export prices, rpx_{ist}, are similarly adjusted for the exchange rate, and measure the price of exports in sector s relative to the price of

[44] All variables are denoted in logged form.

the same sector's output in each trading partner, weighted according to the share of exports of s that go to the partner country. The trade-weighted exchange rate is constructed in the same way as the relative import price above, except that bilateral export data are used to construct the country weights.

- Domestic Income, yd_{ist}, measures the whole-economy GVA for country i. This is a more suitable choice of income measure than sector-only income, since the demand for imports of a particular sector (say, k) is not confined solely to demand from the *domestic* k sector.[45] As a result, the yd_{ist} terms are common to all S sectors in each country: $\mathsf{yd}_{ist} = \mathsf{yd}_{it}$ for all $s = 1, \ldots, P$. This does not pose any problem for empirical modeling, although the CCE estimator in country-by-country analysis does not include the mean of domestic income terms, since they do not vary across individual sectors.

- Foreign Income, yf_{ist}, measures the weighted whole-economy GVA for each partner country, where countries are weighted by the share of sector s's exports they demand.

[45] For example, whilst the domestic Electrical Equipment sector might import goods from foreign Electrical Equipment sectors, the import measure for that category includes demand for Electrical Equipment imports from other sectors, such as Wood Products.

2.B Additional results

2.B.1 Unit root and cointegration test statistics

The test statistics from unit root tests on the level of all six variables of interest are presented in Table 2.11, whilst cointegration test statistics are reported in Table 2.12.

Table 2.11: PANEL UNIT ROOT TESTS

Variable	Lag Order	CADF$^{\diamond}$	BD	IPS$^{\diamond}$
Exports‡	1	-1.936^{**}	0.458^{**}	-2.150^{**}
	2	-1.330^{**}	0.270^{**}	-1.761^{**}
	3		-0.761^{**}	
Imports‡	1	-2.568^{**}	-0.500^{**}	-2.708
	2	-1.845^{**}	-0.407^{**}	-2.735
	3		-0.467^{**}	
Relative Export Price†	1	-1.810^{**}	-1.047^{**}	-1.561^{**}
	2	-1.882^{**}	0.538^{**}	-1.197^{**}
	3		0.080^{**}	
Relative Import Price†	1	-1.137^{**}	0.151^{**}	-1.167^{**}
	2	-0.721^{**}	0.530^{**}	-0.965^{**}
	3		0.342^{**}	
Foreign Income‡	1	-3.887	-0.072^{**}	-2.803
	2	-2.261^{**}	-0.289^{**}	-2.535
	3		-0.129^{**}	
Domestic Income‡	1	-2.560^{**}	-0.145^{**}	-3.445
	2	-1.283^{**}	-0.153^{**}	-2.711
	3		-0.266^{**}	

NOTES: Tests, performed on the full sample of $NS = 143$ individuals over $T = 15$ time periods, are as follows: CADF is the cross-section augmented DF test of Pesaran (2007); BD is the cross-section robust test of Breitung and Das (2005); IPS is the earlier test of Im *et al.* (2003), which assumes cross-section independence. Symbols are as follows: $\diamond$ indicates that truncated test statistics have been used, where appropriate, to account for the small T property of the sample; † indicates that tests were performed assuming the presence of a constant alone; ‡ indicates both a constant and linear time trend were included. ** denotes a failure to reject the null hypothesis (of a unit root) at the 1% significance level, for which critical values are: CADF: †: -2.29; ‡: -2.93; BD (asymptotic): † and ‡: -1.945; IPS: †: -1.75, ‡: -2.42. *GAUSS code:* `ub_robust.src` (BD); `CIPSmarch06.prc` (CADF); `IPSmarch06.prc` (IPS). BD code was kindly provided by Joerg Breitung, and CADF and IPS code by M. Hashem Pesaran.

Table 2.12: PANEL COINTEGRATION TESTS

		Imports	Exports
H_{NOHET}	Pooled sample	-4.888	-2.733
H_C: Country by Country tests	Belgium	-1.428^{**}	-1.884^{**}
	Canada	-0.206^{**}	-1.798^{**}
	Denmark	-2.305	-1.751^{**}
	Finland	-3.230	-1.578^{**}
	Germany	-1.088^{**}	-2.853
	Italy	-3.202	0.441^{**}
	Japan	1.504^{**}	-2.141
	Netherlands	-1.537^{**}	-1.583^{**}
	Norway	1.210^{**}	1.745^{**}
	UK	-2.364	1.188^{**}
	US	-1.231^{**}	0.404^{**}
H_I: Industry by Industry tests	Food, beverages and tobacco	-1.940^{**}	-1.227^{**}
	Textiles, leather and footwear	-1.918^{**}	-0.606^{**}
	Wood and wood products	-1.525^{**}	0.479^{**}
	Paper, printing and publishing	-0.390^{**}	-0.337^{**}
	Chemical products	-1.855^{**}	-1.614^{**}
	Rubber and plastics	-2.535	-0.912^{**}
	Non-metallic minerals	-1.963	0.704^{**}
	Basic metal products	-0.813^{**}	-1.094^{**}
	Machinery and equipment	-1.126^{**}	-1.446^{**}
	Other machinery and equipment	-1.410^{**}	-0.979^{**}
	Electrical and optical equipment	-0.747^{**}	-1.254^{**}
	Transport	-1.594^{**}	-0.749^{**}
	Other Manufacturing	-1.502^{**}	-1.113^{**}

NOTES: Figures are test statistics from a test for one cointegrating vector (against an alternative of none); ** indicates that the null hypothesis is accepted at the 1% significance level (critical value: 1.946). *GAUSS code:* `pan2step-ctest.src`, written by the author and including original code from Joerg Breitung to obtain long-run parameter estimates.

Table 2.13: LONG-RUN ELASTICITIES: POOLED EQUILIBRIUM-CORRECTION RESULTS

Sample	Exports (x)		Imports (m)	
	Price (rpx)	Income (yf)	Price (rpm)	Income (yd)
Pooled sample	−1.346	1.232	−0.847	1.403
	(0.120)	(0.148)	(0.123)	(0.153)
AR(1): F(1, 1703)	9.650 [0.0019]		12.086 [0.001]	
AR(2): F(2, 1702)	21.791 [0.000]		11.330 [0.000]	

NOTES: Estimation period is 1988-2002 using annual data. Estimates obtained via a pooled OLS regression including variables in first differences and lagged levels, as discussed in the text, with long-run coefficients normalised on exports (x) or imports (m) as appropriate. Figures in (·) denote standard errors. In the case of complete pooling, data have been pooled across all $NS = 143$ country-sector individual units, for $T = 15$ time periods. *Ox code:* `TradeEstimation.ox`.

2.B.2 Pooled Equilibrim-Correction estimation

The Breitung two-step estimates for the common long-run elasticities in Table 2.3 can be compared to conventional elasticities obtained from a pooled equilibrium-correction model, reported in Table 2.13. This involves running each trade regression with the trade flow as the explanatory variables, and the lagged levels and first differences (with appropriate lags of the latter) of the trade flow, relative price, and income series. The results are reasonably close to the point estimates from the Breitung procedure applied to the completely pooled sample, but as discussed above, the homogeneity assumption embodied in this approach is not justified by the evidence.

Part 2

ROBUST FORECASTING STRATEGIES:

THEORY AND APPLICATIONS

Chapter 3

Forecasts, Densities and Decisions: A View from the Literature

Tu ne prévois les événements que lorsqu'ils sont déjà arrivés

EUGÈNE IONESCO
Le Rhinocéros, act 3

What kind of questions does an econometrician face when trying to forecast a stochastic process such as inflation or output growth? The most obvious one is why the forecast needs to be produced at all, and in many cases the answer is that the prediction is an intermediate input in a wider decision-making process. Having established the need for the forecast, the next question might ask what the characteristics are of the underlying process, since an approach that might be suitable for a series that varies little over time, might be of limited use if there are large unpredictable fluctuations. Following on from this is the issue of how an appropriate forecast might be provided, and whether it is important to provide a prediction only of the most likely value of the process, or a more comprehensive expression of the *uncertainty* surrounding this point forecast. Finally, an important exercise *ex post* is to evaluate how good the forecasts actually were.

From these issues emerge a number of points that encapsulate the approach to forecasting models in this thesis, which can be summarised as follows:

- Density forecasts are important, for two reasons. First, they provide a more comprehensive expression of forecast uncertainty than a point expectation alone, and secondly, they demonstrate how forecasts are relevant in decision-making environments.

- Some of the biggest challenges facing a forecaster come from processes that change over time, due to unpredictable events that lead to breaks in their elements (such as the mean, or trend growth rate).

- In order to avoid systematic forecast failure, a forecast model must be capable of adapting to such phenomena, either directly (where a break process is modeled explicitly), or indirectly (where a more 'mechanistic' correction method is used).

- Finally, in addition to the need for robustness in forecasting models is a requirement that methods used to *evaluate* forecast accuracy are also robust to breaks.

To see how these issues are relevant, the remaining parts of this chapter set out the context and terminology behind the forecasting models used later in the thesis, emphasising the impact of structural breaks, and the role of forecasts in decision making.

3.1 Forecasting with Structural Breaks

All of the forecasting exercises in this thesis share a common set of concepts and terminology and, since notation is not used consistently across the literature, it is helpful to start this chapter by explaining the basic motivation for making a forecast.[1]

In general, a forecast is made at a discrete point in time T, which is known as the forecast origin. The period over which it relates to is the forecast horizon, whose length is denoted h; thus at time T, a projection for $T+h$ is an h-step ahead forecast.

The variable being forecast is an n-dimensional stochastic process $\mathbf{Y}_t = (Y_{1,t}, \ldots, Y_{n,t})'$, whose joint density function is given by the data generating

[1]The notation here builds on that developed by Clements and Hendry (1998).

process (DGP):[2]

$$\mathsf{f}_{\mathbf{Y}_t}(\mathbf{y}_t|\Omega_{t-1},\theta_t) \overset{\mathsf{D}}{=} \mathsf{D}_{\mathbf{Y}_t}\left(\mathbf{y}_t|\Omega_{t-1},\theta_t\right),$$

where the k-dimensional vector of parameters $\theta_t \in \mathbf{\Theta} \subseteq \mathbb{R}^k$, for all t. The information set Ω_t includes any necessary pre-sample information $\mathbf{Y}_0$, and the history of the $\mathbf{Y}_t$ process $\mathbf{Y}_{t-1}^1 = (\mathbf{y}_1 \dots \mathbf{y}_{t-1})$, so $\Omega_{t-1} = (\mathbf{Y}_0 : \mathbf{Y}_{t-1}^1)$; further, the t subscript in the DGP allows the true density to change over time. A sequence of realisations of $\mathbf{Y}_t$ gives the actual values of the process over a specific range: thus $\{\mathbf{y}_t\}_{t=1}^n$ denotes realisations (in lower case) from $t = 1$ to n.

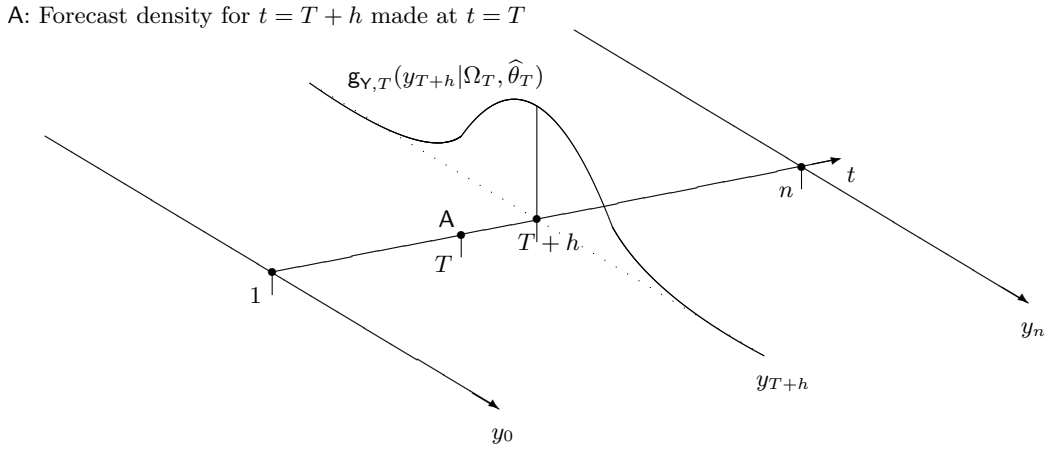
Since there is no reason to presume that $\mathsf{f}_{\mathbf{Y}_t}(y_t)$ is known, the task in hand is to produce a forecast of the density of Y. Using the forecast origin T from above, an h-step forecast of $\mathsf{f}_{\mathbf{Y}}(\cdot)$ is written as $\mathsf{g}_{\mathbf{Y},T}(y_{T+h})$. Since this function might make use of estimated parameters at T, and the observable history of Y up to this point, a more complete specification of the forecast is $\mathsf{g}_{\mathbf{Y},T}(y_{T+h}|\Omega_T, \widehat{\theta}_T)$. This general formulation highlights the fact that estimation might be required in order to obtain $\widehat{\theta}_T$, and subsequent chapters discuss the role of such estimates in forecasting exercises.

The intuition behind a density forecast is aided by a graphical example, and so Figure 3.1 demonstrates a possible density of a univariate process Y_t at time $t = T$. Rather than forecasting a point expectation, the complete probability distribution function is forecast; this produces an explicit probability of every conceivable outturn of Y_T, rather than the mean outturn. Thus a density forecast can offer a more comprehensive specification of the uncertainty surrounding a forecast than might be possible with a point or interval forecast.

Precisely how this density forecast is produced will vary with the context; for example, the Survey of Professional Forecasters (see Clements (2005, Chapter 5), or Diebold, Hahn and Tay (1999)) produces a discrete probabil-

[2] To avoid confusion, $\mathsf{f}_{\mathbf{Y}_t}(\cdot)$ represents the data generating process in subsequent chapters; $\mathsf{D}_{\mathbf{Y}_t}(\cdot)$, on the other hand, denotes a general density function; the Nomenclature on page xii provides a glossary of key terms.

Figure 3.1: DENSITY FORECASTING



ity distribution of the inflation and output growth outturns over a forecast horizon, based on survey responses. In contrast, the National Institute of Economic and Social Research produces a forecast based on projections from its economic model (see Poulizac, Weale and Young (1996)), whilst the Bank of England's 'fan chart' of inflation is based on the inflation projections produced by its economic model, and decided upon by the Monetary Policy Committee (see Chapter 7). What is clear is that a density forecast may come from a variety of sources, encompassing survey-based methods and explicit econometric models, amongst other things.

Further considerations relate to the specification of a forecast density. In some cases, a functional form is imposed (such as a two-piece Gaussian likelihood for the Bank of England forecasts: see Britton, Fisher and Whitley (1998)). In this case, the density forecast is given by the relevant parameters characterising the density function (such as the mean and variance for a normal distribution). For more general specifications, forecast densities are estimated using kernel methods (see, for example, Pagan and Ullah (1999), or Silverman (1986)); this method is used for exchange rate forecasts by Sarno and Valente (2004).

3.1.1 Predictability and the role of information

An important statistical property behind forecasting exercises is the predictability of a process, which is distinct from a researcher's ability to forecast it. Predictability in this sense concerns the probability density of a random variable, and revolves around whether it is affected by past events. More concretely, the random variable Y_t is unpredictable with respect to its information set Ω (which includes as a minimum, the sigma-field generated by its past), if its density function $\mathsf{D}_{\mathsf{Y}_t}(\cdot)$ satisfies the condition:

$$\mathsf{D}_{\mathsf{Y}_t}(y_t|\Omega_{t-1}) = \mathsf{D}_{\mathsf{Y}_t}(y_t).$$

This formulation highlights the parallel with statistical independence, where for random variables A and B, if $\mathsf{P}(A|B) = \mathsf{P}(A)$ then the two are independent.

An immediate implication of *un*predictability is that knowing the past history of a process, and *all* possible relevant information, cannot improve the prediction of any future outturns of Y_t, and will not reduce any uncertainty about the process. Thus it follows immediately that unpredictability precludes the ability to forecast, in the sense that a 'forecast' of an unpredictable process has no relevant information to exploit when making a prediction. In the context of structural change it is commonly the break itself which is unpredictable, even if it enters into a process whose other components may be predictable: an example to illustrate this is provided in Chapter 6.

Secondly, understanding the role of information in forecasting is critical to the success of different approaches: predictability is a necessary condition, but by no means sufficient. Knowing the processes that could generate Ω_{t-1} is particularly valuable, since in practice there is no one-to-one mapping from a set of events and a particular information set. Thus the ability to distinguish between an upturn in inflation due to a demand-side increase in income, and a supply-side increase in oil prices, for example, may be critical in the success of forecasting the next outturn of inflation. Further,

knowing *how* Ω_{t-1} enters into the density function of a process is equally important, and in this regard a flexible non-linear approach to modelling and forecasting can be particularly valuable. In the case of a scalar linear model, such as

$$y_t = \boldsymbol{\beta}_t' \mathbf{x}_{t-1} + \epsilon_t \quad \text{where } \epsilon_t \sim \mathsf{IN}(0, \sigma_\epsilon^2) \tag{3.1}$$

forecast failure could be induced by a change in the behaviour of $\mathbf{x}$, or by a change in the parameter $\boldsymbol{\beta}$. It is breaks in the latter – where, for a *given* level of $\mathbf{x}_{t-1}$ the corresponding y_t differs – that are explored in the empirical application to money demand foreasts in Chapter 6.

3.1.2 The impact of structural breaks

Of the problems encountered when trying to build a forecast model, this thesis focuses on the impact of structural breaks in the elements of a data generating process. Shifts in deterministic elements form an important part of a wider taxonomy of sources of forecast error developed by Clements and Hendry (1998, 1999). Within this framework, equilibrium-mean changes, slope shifts, mis-specification and forecast origin uncertainty all play a significant role in determining forecast errors, and a delineation of sources is helpful in understanding which sources of forecast errors might be the most pernicious.[3]

The formal taxonomy is not established here: the interested reader can consult the references cited above for further details; however, the issue is relevant to the work in this thesis, since it provides a context for the study of location shifts as a critical source of systematic forecast failure in some forecast models. Further, it also provides a direction for the exploration of robust forecast devices – that is, models that do not entail systematic forecast failure whenever there is a structural break. Chapter 4 elaborates on this issue, providing a more concrete conception of a robust device.

In general, shifts may be caused by political issues (such as the September

[3] Clements and Hendry (2002, p. 543), amongst other sources, derives the error taxonomy.

2001 terrorist attacks in America, or the recent war in Iraq), legal matters (such as the 1984 Banking Act, as discussed in Chapter 6) or technological changes (such as the introduction of digital television): an overview is presented in Clements and Hendry (2006). Irrespective of their causes, breaks can pose a serious problem to many kinds of forecasting exercises.

To demonstrate a structural break in an economic model, and its potential effects, a simple example is useful. A univariate random variable Y_t has a DGP given by a simple equilibrium-correction mechanism:

$$y_t = \begin{cases} \alpha_1 + \beta_1 y_{t-1} + \epsilon_t & t < T \\ \alpha_2 + \beta_2 y_{t-1} + \epsilon_t & t \geq T \end{cases}, \tag{3.2}$$

where $|\beta_i| < 1$ for $i = 1, 2$ and $\epsilon_t \sim \mathsf{NID}\left(0, \sigma_\epsilon^2\right)$. This means to say that the break in the DGP relates to the intercept parameters α_1 and α_2, and the autoregressive persistence parameters β_1 and β_2. It is possible for the error variance σ_ϵ^2 to change (as it can, for example, in Pesaran and Timmermann (2004)), but in this example, it remains constant.

Up to the breakpoint at T, and given assumptions about the initial value of y_0,[4] the unconditional distribution of y_t is given by

$$y_t \sim \mathsf{N}\left(\frac{\alpha_1}{1-\beta_1}, \frac{\sigma_\epsilon^2}{1-\beta_1^2}\right).$$

After the break, however, the process has a new unconditional expectation $\mathsf{E}[y_t] = \frac{\alpha_2}{1-\beta_2}$, which could be different from the original expectation, with potentially serious implications for the forecast performance of a model based only on pre-break information. A critical element in the behaviour of Y_t after a break is therefore the value of α_i: if $\boldsymbol{\alpha} = (\alpha_1, \alpha_2)' = \mathbf{0}$, and $\beta_1 \neq \beta_2$, then the structural shift only affects the unconditional variance, but not the mean. Thus forecasts will have a different (unconditional) variance,[5] but will remain unbiased.

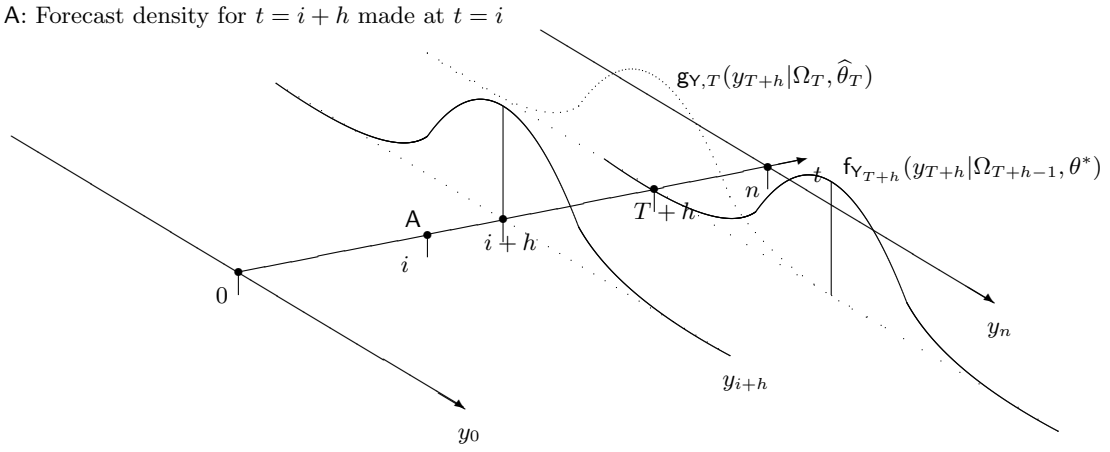
If α changes, but $\beta_1 = \beta_2$, then there is a location shift in the mean of Y_t,

[4] Namely, that

$$y_0 \sim \mathsf{N}\left(\frac{\alpha_1}{1-\beta_1}, \frac{\sigma_\epsilon^2}{1-\beta_1^2}\right).$$

[5] Even though the *conditional* variance remains σ_ϵ^2.

Figure 3.2: LOCATION SHIFTS IN THE FORECAST DENSITY



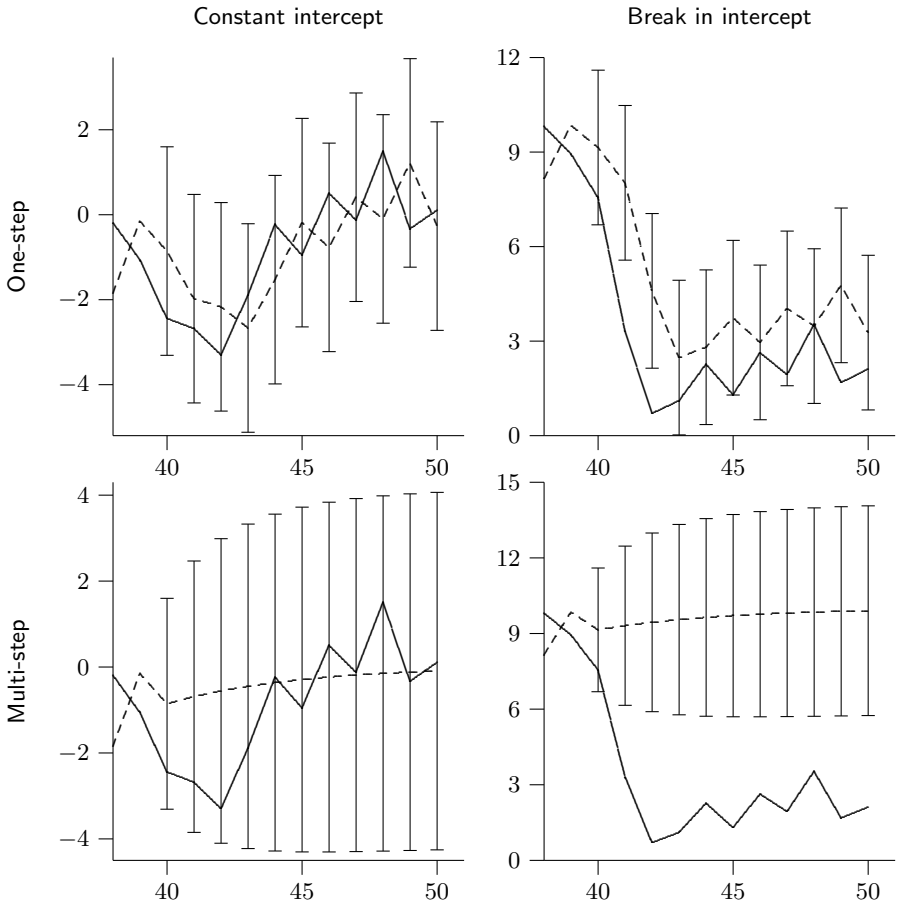
but the variance remains constant. In this case, the post-break parameter vector θ is denoted θ^*; if the *forecast* model is still based on θ, then forecasts will be biased, since $\mathsf{E}[Y_{T+h}] \neq \mathsf{E}[\widehat{Y}_{T+h|T}]$, where $\widehat{Y}_{T+h|T}$ denotes the h-step ahead forecast of Y_{T+h}. This is demonstrated graphically in Figure 3.2; even though the true density of Y shifts at T, the forecast density (indicated by a dotted line) does not. Thus there is a discrepancy between the two densities; Chapter 4 addresses ways in which this discrepancy can be quantified, and then the discrepancies of various forecast models compared.

The effect of a structural break when $\boldsymbol{\alpha} = \mathbf{0}$ compared to $\boldsymbol{\alpha} \neq \mathbf{0}$ is demonstrated by a simple example in Figure 3.3. The DGP in (3.2) is simulated for a sample size of 50, with a break at $T = 40$, for parameters $\beta_1 = 0.9$, $\beta_2 = 0.5$, $\sigma_\epsilon^2 = 1$ and on the left hand panels, $\alpha_1 = \alpha_2 = 0$, and on the right, $\alpha_1 = \alpha_2 = 1$. Using in-sample estimates of $\widehat{\alpha}$ and $\widehat{\beta}$ (on the assumption that an econometrician would be unaware of the break), out-of-sample forecasts for the post-break period can then be produced.

Addressing the case with a constant intercept (at zero), the 1-step and multi-step forecasts are reasonably good, since none falls outside of the 95% standard error bars.[6] Thus it does not seem as though there is major forecast

[6]Multi-step refers to a fixed forecast origin – in this case, the break point T – and

Figure 3.3: EQUILIBRIUM-CORRECTION FORECASTS WITH BREAKS



NOTES: These graphs represent forecasts produced from a simulated AR(1) process, as described in the text, which has constant parameters during the sample estimation period $t = 1, \ldots, 40$, but then undergoes a break in the autoregressive parameter ϕ at $t = 40$, from $\phi = 0.9$ to $\phi = 0.5$. On the left-hand column are the 1-step and multi-step forecasts when there is no intercept in the data generating process, and on the right-hand column, when there is an intercept of unity for all $t = 1, \ldots, 50$.

failure, and since the mean of both forecast and true densities of Y_t is zero, so the forecasts are unbiased.

In contrast, the same simulations when there *is* a break in the intercept leads to dramatic deterioration of forecast performance. Three of the 1-step forecasts fall outside of the 95% bands, and all appear to have bias, since they are consistently below the realised value of Y_t. Similarly, the multi-step

h-step forecasts as h grows from $h = 1$ to $h = 10$.

forecasts are poor, with all error bars beyond the 1-step case failing to cover the outturn of Y_t. The reason for this lies in the fact that the forecasts are based on estimation of the in-sample model, when the unconditional mean of Y_t was $\mathsf{E}[Y_t] = (1 - 0.9)^{-1} = 10$; in contrast, the post-break mean is $\mathsf{E}[Y_t] = (1 - 0.5)^{-1} = 2$. The forecast model consistently predicts that Y_t will try to correct back to the *pre*-break equilibrium, whereas in reality, correction takes place towards the *post*-break mean.

This simple demonstration raises a number of questions. Pesaran and Timmermann (2004, 2005) consider the problem of using pre- and post-break sample data, with a view to deriving an optimal 'window length' of each, to obtain estimators of α and β that produce better forecasts (when measured by mean-squared forecast errors) than a simple case when only (say) pre-sample data are used. Their work focusses on finite-sample estimation of parameters in the presence of structural breaks, which is useful to forecasting exercises when sample sizes are very small (i.e. less than 50), which is common for many macroeconomic examples (such as the forecast evaluation in Chapter 7). However, the relevance with regard to the forecasting situations considered in this thesis is limited by the assumption that the unconditional mean does not change (Pesaran and Timmermann (2005, Proposition 2)); in contrast, one of the fundamental themes underlying this, and later chapters, is the effect of breaks, and in particular, location shifts, on methods of density evaluation, and on a range of forecasting models. Thus the analysis here, follows in the spirit of Clements and Hendry (1998, 1999), in emphasising the importance of structural breaks in a forecasting exercise *when the mean changes*, and the pressing need to develop methods that are robust to such breaks – that is, forecasts that are not permanently or irrevocably 'damaged' by a shift; Chapter 4 discusses this issue in greater depth.

3.2 Why are density forecasts important?

As Timmermann (2000) comments in an introduction to a special issue of the *Journal of Forecasting* on density forecasting:[7]

[7] Timmermann (2000, p.234).

> 'Density forecasting is fast becoming an important tool for decision makers in situations where loss functions are asymmetric and forecast errors follow non-Gaussian distributions.'

The motivation for studying densities, rather than point predictions, stems from an interest in the role of forecasts as intermediate inputs in a wider decision-making environment. For example, the Bank of England (like other central banks) uses density forecasts in order to make decisions about the path of interest rates that is most consistent with achieving a target rate of inflation over a given horizon. The practice of using forecasts in this kind of environment is part of a wider branch of decision theory and forecasting. Chamberlain (2000) discusses the relationship between decision theory and econometrics, emphasising the need for distributional forecasts, and this concern underpins the density forecasting and evaluation literature, which is dominated by two sets of authors. On one hand, Granger, Pesaran and others[8] draws on game theory to evaluate forecasts in an economic decision-making context. An alternative approach, emphasising statistical evaluation of forecasts, is developed by Diebold and co-authors.[9] Since the methods of forecast evaluation employed in Chapter 7 follow the latter, the decision-theoretic context of density forecasts is explained below using the results derived in Diebold *et al.* (1998).

The starting point for the decision problem facing an agent is a cost (or loss) function that provides a complete specification of the decision problem across all possible actions. This produces a profile of the costs associated with each action, and leads naturally to the decision problem, which in this case is to choose an action to minimise the expected loss. In this model, the loss function is given by $L(a_t, y_t)$, where a_t refers to an agent's action choice, and y_t is the value of a stochastic variable that affects her costs, and therefore her decisions. Further, the true density function of Y_t is given by $\mathsf{f}_{\mathsf{Y}_t}(y_t|\Omega_{t-1}, \theta)$, as discussed in Section 3.1; since this data generating

[8] See Granger and Pesaran (2000*a*, 2000*b*), Pesaran and Skouras (2002) and Granger and Machina (2006).

[9] See Diebold, Gunther and Tay (1998), Diebold, Hahn and Tay (1999), Diebold, Tay and Wallis (1999) and Tay and Wallis (2002).

process is unknown, the agent makes a forecast of the density of Y_t, which is denoted, as before, by $\mathsf{g}_{\mathsf{Y},t-h}(y_t|\Omega_{t-h},\theta_{t-h})$, which she believes to be a congruent representation of the DGP. Then, given this density forecast, the optimal decision a_t^* is chosen to satisfy:[10]

$$a_t^*(\mathsf{g}_{\mathsf{Y},t-h}(y_t|\cdot)) = \underset{a_t\in A_t}{\arg\min} \int_{-\infty}^{\infty} L(a_t,y_t)\mathsf{g}_{\mathsf{Y},t-h}(y_t|\Omega_{t-h},\theta_{t-h})\ dy. \tag{3.3}$$

For every realization of Y_t, the action choice a_t^* defines the loss $L(\cdot)$ faced. A crucial feature of the optimal decision is that it is a function of the forecast density: thus the *quality* of the forecast is important. Equation (3.3) therefore gives the optimal decision that is 'robust', in a sense, to the forecast outturns of Y_t: a_t^* minimises the loss across *all* forecast realisations of Y_t, with the weight attached to the loss for a particular realisation y_t given by the density forecast.

The loss function itself is a random variable, and possesses a probability distribution which Diebold *et al.* (1998) term the loss distribution, and which depends only on the action choice. Further, the *realised loss* – that is, the loss incurred from an action profile a_t^*, when weighted by the *true* density of Y_t, has expectation:

$$\mathsf{E}\left[L(a_t^*,y_t)\right] = \int_{-\infty}^{\infty} L(a_t^*,y_t)\mathsf{f}_{\mathsf{Y},t}(y_t|\Omega_{t-1},\theta)\ dy. \tag{3.4}$$

In summary, the user's density forecast determines the optimal action choice a_t^*; and in turn, this yields a distribution of the actual losses faced by the user, for any conceivable realisation y_t. As Diebold *et al.* (1998) comment, '[a] density forecast translates into a loss distribution.'[11] This gives rise to three remarks.

First, it is impossible to derive a ranking of two *incorrect* density forecasts such that all users would agree with, *irrespective of their loss functions.* As Diebold *et al.* (1998, Proposition 1) demonstrates, two incorrect density functions cannot be ranked according to expected loss, since such a rank-

[10] Following Diebold *et al.* (1998), a unique minimiser is assumed, for which sufficient conditions are that A_t is a compact set, and $L(\cdot)$ is strictly convex in a_t.

[11] p. 6.

ing would not be invariant to the specification of the loss function: relating back to Arrow's impossibility theorem, 'the ranking effectively reflects a social welfare function, which does not exist.'[12]

Secondly, the complete loss function specification and cross-sectional aggregation rules which Pesaran and Skouras observe are necessary in a decision-based approach to forecast evaluation are, in most real-world situations (such as monetary policy setting), requirements which cannot be met: an agent's loss function may be unknowable, and simply cannot be aggregated across all agents.

Thirdly, in the context of decision making when the actions of the decision maker may affect the forecast process (e.g. where a central bank forecasts inflation), these problems are compounded.

However, these problems are not insuperable. As Diebold *et al.* (1998, Proposition 2) demonstrates, the correct density forecast is weakly superior to all other forecasts, *irrespective* of agents' loss functions. The reason for this is that when the forecast is correct, so $\mathsf{f}(\cdot) = \mathsf{g}(\cdot)$, (3.3) implies that the action choice a_t^* in reality minimises (3.4); thus for any other density forecast k, the corresponding action choice $a_{k,t}^*$ cannot have a lower expected loss. A crucial implication of this result is the fact that the best (loss minimising) forecast is a correct one – and so forecast evaluation should be directed towards assessing whether forecast densities are correct. This result underpins many of the forecast evaluation exercises in Diebold *et al.* (1998) and Clements (2005), amongst others, and provides a framework for the evaluation exercises in Chapter 7.

3.3 Conclusion and directions

In summary, the examination of the forecasting process in this thesis focuses on three main aspects. First, a primary motivation is the context within which forecasts are made: by acknowledging that forecasts, whilst having intrinsic value in their own right, are also important intermediate inputs into

[12] *op. cit.*, p. 7.

a wider decision-making environment, it is possible to draw on an expanding literature on *density* forecasting in decision theory.

Following on from this, the effects of structural breaks on the forecasting process – including the production and evaluation of forecast models – can be significant, rendering some forecasts permanently incorrect, and some evaluation methods unreliable. In the light of this, a deeper understanding of how and why models and evaluation tests might go wrong is important.

This raises a fundamental issue from the practical point of view of forecast model selection in the future: a forecast that may have performed well in-sample could fail completely out of sample as a result of a break, and so an examination of the effects of breaks must also be motivated by a desire to build a *robust* forecasting device, that is not systematically affected by a break.

Chapter 4

Robust Forecasting Models

> This is one time where television really fails to capture the true excitement of a large squirrel predicting the weather
>
> BILL MURRAY *as* PHIL CONNORS
> *Groundhog Day*

In the absence of a crystal ball, or other exotic devices, unexpected breaks can have severe effects for unwary forecasters. If models are particularly vulnerable to breaks, then location shifts in the unconditional mean of the process being forecast can induce systematic forecast failure, so that forecasts persistently may over- or under-shoot the actual realizations of the process. Further, not only may bias exist in the short run, but it may remain asymptotically.

Quantifying forecast failure has been the subject of a number of studies, as discussed by Granger (2001) or Clements and Hendry (1998, Chapter 3), amongst others. The approach taken in this chapter continues within the framework of forecast *densities*: thus metrics for evaluating robustness are introduced for complete density functions, rather than point expectations alone. As a result, tools developed to quantify the 'global distance' between two density function, such as integrated errors, or relative entropies, can be applied to forecast models: the relative entropy of the forecast errors of a particular model provides a measure of the 'cost' of using that model, by evaluating its forecast error density in relation to the density of the innovation term in the underlying data generating process that is being forecast. Since the cost of using a forecast model may change in response to structural breaks or other factors, a 'robust' model is one whose cost *does not increase permanently* after a structural break.

Defining robustness in this way is useful for a number of reasons. First, it leads naturally to the evaluation of models using a class of metrics that are suited to error densities, which are of interest in this thesis.

Following on from this, it provides a framework to evaluate, via complete densities, existing model transformations and robust devices that have hitherto concentrated solely on comparisons of the first and second moments of forecast errors.

Chapter 3 identified equilibrium-correction mechanisms as a class of models that are vulnerable to location shifts. Thus they form the theoretical basis of the transformations, or devices, studied in this chapter, which seek to attenuate the effects of structural breaks. In this sense, there is a focus on 'robustness': the objective is to explore various transformations and find those that make equilibrium-correction models most robust to location shifts, so that breaks do *not* induce systematic forecast failure. Predictably, making a model robust comes at a cost; thus the chapter also examines ways to quantify this cost, and the factors affecting it, with a view to choosing the 'least-cost' device for a given forecasting exercise.

4.1 Robustness and Density Error Metrics

In order to evaluate the error density of a forecast model, a metric or measure of distance is necessary, to compare it to the density of the underlying innovation that forms a part of the data generating process. The fact that the DGP does comprise a random innovation element implies that no forecast model can outperform this irreducible component; therefore the DGP innovation error provides a benchmark against which the forecast models can be evaluated.

Within a general specification, the innovation density function is denoted $f(x)$, and the forecast error density is given by $g(x|\boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is a vector of parameters that characterise the error density. For notational simplicity below, conditioning on the parameter vector is left implict, but should be borne in mind. A metric of 'global distance' therefore measures the distance between the density functions. In the best case scenario from the perspective

of forecast model performance, the densities are identical, so $g(x) = f(x)$, $\forall x$, in which case the distance between the two is zero. In all other cases, there will be a non-zero distance at some point between the two densities. This suggests that an appropriate metric should have the same properties.

Fortunately, as Ullah (1996), or Pagan and Ullah (1999, Chapter 2) discuss, a number of suitable candidate distance measures exist, and those considered here fall into two categories. The first involves taking a function of the difference between two densities at a particular point on their support, and then integrating across the whole support.[1] For instance, one common measure of global distance in this vein is the integrated squared error (ISE) between $f(\cdot)$ and $g(\cdot)$, denoted I_{SE} and given by:

$$I_{SE}(f,g) = \int_{\mathcal{X}} (f(x) - g(x))^2 \, dx, \tag{4.1}$$

where $\mathcal{X}$ is the range of integration, and $x \in \mathcal{X}$.[2]

An extension of this is found by taking expectations with respect to $f(x)$, which yields the mean integrated squared error:

$$MISE = \mathsf{E}[I_{SE}] = \int_{\mathcal{X}} (f(x) - g(x))^2 \, f(x) \, dx.$$

Both the ISE and MISE have the properties that they are non-negative, and equal to zero if and only if $f(x) = g(x)$; further, the ISE is symmetric, so that $I_{SE}(f,g) = I_{SE}(g,f)$. However, neither count as metrics in a strict sense, since they both violate the triangle inequality; therefore it is with some abuse of terminology that these measures are referred to below as distances.

A second class of discrepancy measures includes the Kullback-Leibler (KL) information distance, or relative entropy. The KL distance is a well-known measure (see, for instance, Silverman (1986), Gallant (1997), Davison

[1] Implicitly, if the densities are identical, then their supports should be equivalent; where they differ, the range of integration must cover both supports. This point is developed below.

[2] Following from the previous footnote, denoting X_i $(i = f, g)$ as the support for density i, then with $\overline{x}_i = \sup\{x_i \ : \ x_i \in X_i\}$ and $\underline{x}_i = \inf\{x_i \ : \ x_i \in X_i\}$ denoting the upper and lower bounds respectively of each support, $x \in [\min\{\underline{x}_f, \underline{x}_g\}, \max\{\overline{x}_f, \overline{x}_g\}]$.

(2003), Pagan and Ullah (1999)[3] or Rao (1973)) that, like the ISE, is not a strict metric, since it violates the triangle inequality and the requirement of symmetry.

The KL distance is denoted $I_{KL}(f, g)$ and is given by:

$$I_{KL}(f, g) = \int_{\mathcal{X}} f(x) \log \left\{ \frac{f(x)}{g(x)} \right\} \, dx. \tag{4.2}$$

As above, $I_{KL} \geq 0$, and $I_{KL} = 0$ iff $f(x) = g(x)$.

Relative entropies and distance measures have recently emerged in the literature on forecast evaluation: Sarno and Valente (2004) and Mitchell and Hall (2005) use the ISE and KL discrepancies respectively, although the focus in each study is different to that of this thesis. Where the former evaluates the distance between two competing density forecasts and an estimated true density function, for the point of view of comparing the *ex post* forecast performance of different models, the latter uses the KL distance in the context of density forecast combination and Bayesian model averaging. Since the substantive conclusions drawn below are invariant to the choice of distance measure, and the KL discrepancy is computationally quicker, all results report the KL distance, rather than the ISE or MISE.

4.1.1 Distance measures and forecast errors

Having shown how to compare densities in theory, it is now possible to apply this to forecast errors. The general framework explained above is applicable to a wide class of density functions; however, for the purpose of the models set out in Section 4.2 and the transformations in Section 4.3, it is useful to move from an abstract representation to a concrete example, since numerical calculations of distances are then possible. Thus the models below assume that all error processes, whether DGP innovations or model errors, are normally distributed.

This simplifying assumption is somewhat restrictive, since a general framework is being used to examine a family of density functions that is

[3]Noting the inconsistency between their equations 2.118 on p.54 and 2.132 on p. 61.

completely characterised by its first two moments, and that has already received a considerable amount of attention in the literature to date. However, given that this approach to forecast evaluation does not seem to have been pursued before now, it is helpful to establish results for cases that have already been analysed, before moving on to more general functions. Further, Section 4.3 discusses a number of insights into existing models that arise when examining robust forecast models within this density framework.

4.1.1.1 Global Distance: An Example

In order to demonstrate the behaviour of the ISE, MISE and Kullback-Leibler measures as the forecast density $g(\cdot)$ changes in relation to the baseline density $f(\cdot)$, a simple example is useful. Consider the information statistic for each measure, when the baseline (DGP) innovation density $f(z) \overset{\mathsf{D}}{=} \mathsf{N}(0,1)$, and the model forecast error density $g(x) \overset{\mathsf{D}}{=} \mathsf{N}(\mu, \sigma^2)$. In relation to (4.1), for example:

$$I_{SE}(f,g) = \int_{-\infty}^{\infty} \left(\phi(x) - \phi\left(\frac{x-\mu}{\sigma} \right) \right)^2 \, dx,$$

and for (4.2):

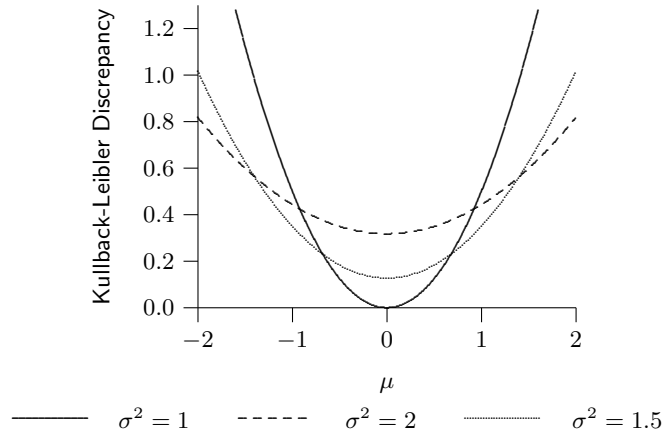
$$I_{KL}(f,g) = \int_{-\infty}^{\infty} \phi(x) \log \left\{ \frac{\phi(x)}{\phi\left(\frac{x-\mu}{\sigma}\right)} \right\} \, dx.$$

where $\phi(\cdot)$, is the density of a standard normal distribution.

On all three measures, the distance between $f(\cdot)$ and $g(\cdot)$ is zero if and only if $f(x) \equiv g(x)$. For deviations in the mean μ, of the model forecast error, from zero, the measures report differing distance statistics, but importantly these are increasing in the (absolute) magnitude of the deviation. Further, the symmetry of each distance measure around $\mu = 0$ reflects the symmetry of a normal density function around its mode.

The role of σ^2 can be demonstrated in Figure 4.1, which shows the Kullback-Leibler discrepancy for three different Gaussian densities, distinguished by their variances. Although it seems at first glance counter-intuitive, for a sufficiently large bias in the mean of each density, the model with the

Figure 4.1: DENSITY BIAS AND VARIANCE EFFECTS



NOTES: The curves show the Kullback-Leibler discrepancy for a normally distributed process $x \sim \mathsf{N}\left(\mu, \sigma^2\right)$, in relation to the baseline process $z \sim \mathsf{N}\left(0, 1\right)$, for three different values of σ^2. *Mathematica code reference:* `Gendistance.nb`

smallest discrepancy is the one with the *largest* variance. The rationale for this is that a biased forecast with a small variance has a large probability mass concentrated around its mean; in contrast, a density with the same bias but a larger variance has less probability mass concentrated around the mean.[4] This could lead to a situation where the distance between the true density and each biased density is smaller for the high variance model than the low variance case. Figure 4.1 suggests that for the solid curve (where $\sigma^2 = 1$), a bias of less than one standard deviation is sufficient for the discrepancy to exceed that of the dashed curve (where $\sigma^2 = 2$).

The fact that a small variance might not be helpful in the presence of bias echoes the famous example of the biased US presidential election prediction published by the *Literary Digest* in 1936, which incorrectly forecast a victory for Alfred Landon against Franklin Roosevelt. The reason for the forecast error lay in the polling method: the poll sample was selected using telephone and car ownership lists, but since these were luxury items, a sample selection bias developed, so that even with a sample of 2,000,000, the estimator was

[4]Where bias is defined as any difference between the mean of the model and the baseline density.

biased and therefore forecast the wrong result.

4.2 Forecasting in Cointegrated Systems

As previous sections have discussed, equilibrium-correction mechanisms, or EqCMs, are particularly vulnerable to location shifts. The intuition behind this is clear: a change in the equilibrium mean of a process implies that any *disequilibrium* in the system will cause the process to 'correct' back to its new mean. If a forecast model *does not capture* the shift in the mean – as might happen if the break occurs after the forecast origin – then it will treat any disequilibrium in the system with reference to the *pre-break* mean, and therefore anticipate a correction towards the *wrong* level, which could lead to the kind of divergence seen in Chapter 3.

In order to relate EqCM-based forecast failure to the evaluation framework introduced in Section 4.1, this section establishes a forecast model based on a cointegrated equilibrium-correction mechanism. Sections 4.2.1 to 4.2.3 develop a multivariate EqCM forecast model, Section 4.2.4 introduces mis-specification, and Section 4.2.5 examines forecast errors in the presence of location shifts.

4.2.1 The data generating process

Moving beyond the $\mathsf{I}(0)$ autoregressive models examined in Chapter 3, this chapter builds on an integrated-cointegrated process, as examined in a wide range of studies.[5] The random variable $\mathbf{x}_t$ is an n-dimenstional $\mathsf{I}(1)$ process, whose in-sample local data generating process (DGP) is given by:

$$\mathbf{x}_t = \boldsymbol{\tau} + \boldsymbol{\Gamma}\mathbf{x}_{t-1} + \boldsymbol{\varepsilon}_t, \tag{4.3}$$

where $\boldsymbol{\varepsilon}_t \sim \mathsf{IN}_n(\mathbf{0}, \boldsymbol{\Omega}_\varepsilon)$ and $\boldsymbol{\tau}$ is an n-dimensional vector of constants. Following Johansen (1995, Chapter 4), $\boldsymbol{\Gamma} = \mathbf{I}_n + \boldsymbol{\alpha}\boldsymbol{\beta}'$, where $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are $n \times r$ matrices of full rank $r < n$, where r represents the number of cointegrating relationships. Thus whilst $\mathbf{x}_t \sim \mathsf{I}(1)$, all the eigenvalues of $\boldsymbol{\Gamma}$ lie on or within

[5] See, for instance, Johansen (1995) and Banerjee, Dolado, Galbraith and Hendry (1993).

the unit circle. Correspondingly, with L denoting a lag operator such that $L^s x_t = x_{t-s}$, no roots of $|\mathbf{I}_n - \mathbf{\Gamma} L| = 0$ lie within the unit circle, $n - r$ roots lie on the circle itself, and r roots lie outside it.

The orthogonal complements $\boldsymbol{\alpha}_\perp$ and $\boldsymbol{\beta}_\perp$ to $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ respectively are $n \times (n - r)$ matrices of full rank, and $\boldsymbol{\alpha}'\boldsymbol{\alpha}_\perp = \boldsymbol{\beta}'\boldsymbol{\beta}_\perp = 0$.

4.2.2 Equilibrium-correction representation

Equation (4.3) can be reparameterised into an equilibrium-correction form, which yields:

$$\Delta \mathbf{x}_t = \boldsymbol{\tau} + \boldsymbol{\alpha}\boldsymbol{\beta}'\mathbf{x}_{t-1} + \boldsymbol{\varepsilon}_t. \tag{4.4}$$

Since $\mathbf{x}_t \sim \mathsf{I}(1)$, it follows that $\Delta \mathbf{x}_t \sim \mathsf{I}(0)$, and $\boldsymbol{\beta}'\mathbf{x}_{t-1}$ yields the $r \times 1$ cointegrating vectors, which are also stationary.[6] As Hendry (2006) notes, both $\Delta \mathbf{x}_t$ and $\boldsymbol{\beta}'\mathbf{x}_{t-1}$ may have non-zero means; to incorporate this, $\boldsymbol{\tau}$ can be written as:

$$\boldsymbol{\tau} = \boldsymbol{\gamma} - \boldsymbol{\alpha}\boldsymbol{\mu},$$

where $\boldsymbol{\gamma}$ is $n \times 1$ and $\boldsymbol{\mu}$ is an $r \times 1$ vector of equilibrium means, so that (4.4) becomes:

$$\Delta \mathbf{x}_t = \boldsymbol{\gamma} - \boldsymbol{\alpha}\boldsymbol{\mu} + \boldsymbol{\alpha}\boldsymbol{\beta}'\mathbf{x}_{t-1} + \boldsymbol{\varepsilon}_t,$$

and so:

$$(\Delta \mathbf{x}_t - \boldsymbol{\gamma}) = \boldsymbol{\alpha}\left(\boldsymbol{\beta}'\mathbf{x}_{t-1} - \boldsymbol{\mu}\right) + \boldsymbol{\varepsilon}_t. \tag{4.5}$$

The expectation of $\Delta \mathbf{x}_t$ and $\boldsymbol{\beta}'\mathbf{x}_{t-1}$ can be established by premultiplying (4.5) by $\boldsymbol{\beta}'$ and taking expectations, yielding:

$$\begin{aligned} \mathsf{E}[\boldsymbol{\beta}'\Delta \mathbf{x}_t] &= \boldsymbol{\beta}'\boldsymbol{\gamma} + \boldsymbol{\beta}'\boldsymbol{\alpha}\mathsf{E}[\boldsymbol{\beta}'\mathbf{x}_{t-1} - \boldsymbol{\mu}] + \boldsymbol{\beta}'\mathsf{E}[\boldsymbol{\varepsilon}_t] \\ &= \mathbf{0}. \end{aligned}$$

Provided it is non-singular, pre-multiplying by $(\boldsymbol{\beta}'\boldsymbol{\alpha})^{-1}$ then yields:

$$(\boldsymbol{\beta}'\boldsymbol{\alpha})^{-1}\boldsymbol{\beta}'\boldsymbol{\gamma} + (\boldsymbol{\beta}'\boldsymbol{\alpha})^{-1}\boldsymbol{\beta}'\boldsymbol{\alpha}\mathsf{E}[\boldsymbol{\beta}'\mathbf{x}_{t-1} - \boldsymbol{\mu}] = \mathbf{0},$$

[6] See, for instance, Banerjee *et al.* (1993, Definition 1, p.84).

and so:

$$\mathsf{E}[\boldsymbol{\beta}'\mathbf{x}_{t-1} - \boldsymbol{\mu}] = -(\boldsymbol{\beta}'\boldsymbol{\alpha})^{-1}\boldsymbol{\beta}'\boldsymbol{\gamma} = \mathbf{0}. \tag{4.6}$$

Thus in equilibrium, $\mathsf{E}[\boldsymbol{\beta}'\mathbf{x}_{t-1}] = \boldsymbol{\mu}$ and when $\boldsymbol{\tau}$ lies in the cointegration space defined by $\boldsymbol{\beta}$, it follows that $\boldsymbol{\gamma} = \mathbf{0}$.

Generally, using (4.5):

$$\begin{aligned}\mathsf{E}[\Delta\mathbf{x}_t] &= \boldsymbol{\gamma} + \boldsymbol{\alpha}\mathsf{E}[\boldsymbol{\beta}'\mathbf{x}_{t-1} - \boldsymbol{\mu}] + \mathsf{E}[\boldsymbol{\varepsilon}_t] \\ &= \boldsymbol{\gamma} - \boldsymbol{\alpha}(\boldsymbol{\beta}'\boldsymbol{\alpha})^{-1}\boldsymbol{\beta}'\boldsymbol{\gamma} \\ &= (\mathbf{I}_n - \boldsymbol{\alpha}(\boldsymbol{\beta}'\boldsymbol{\alpha})^{-1}\boldsymbol{\beta}')\boldsymbol{\gamma} = \mathbf{K}\boldsymbol{\gamma}\end{aligned}$$

where $\mathbf{K}$ is an $n \times n$ non-symmetric idempotent matrix. From Hendry (*op. cit.*), $\boldsymbol{\beta}'\mathbf{K} = \mathbf{0}'$ and $\mathbf{K}\boldsymbol{\alpha} = \mathbf{0}$, which implies that $\mathbf{K}\boldsymbol{\gamma} = \mathbf{K}\boldsymbol{\tau}$. Imposing a constant long-run equilibrium mean yields:

$$\mathsf{E}[\boldsymbol{\beta}'\mathbf{x}_t] = \boldsymbol{\mu},$$

and then from (4.6), allowing $\boldsymbol{\gamma} \neq \mathbf{0}$ implies that $\boldsymbol{\beta}'\boldsymbol{\gamma} = \mathbf{0}$, and so $\boldsymbol{\Gamma}\boldsymbol{\gamma} = \boldsymbol{\gamma}$ and $\mathbf{K}\boldsymbol{\gamma} = \boldsymbol{\gamma}$. Collecting results gives:

$$\begin{aligned}\mathsf{E}[\Delta\mathbf{x}_t] &= \boldsymbol{\gamma} + \boldsymbol{\alpha}\mathsf{E}[\boldsymbol{\beta}'\mathbf{x}_{t-1} - \boldsymbol{\mu}] + \mathsf{E}[\boldsymbol{\varepsilon}_t] \\ &= \boldsymbol{\gamma},\end{aligned}$$

and so (4.5) expresses $\Delta\mathbf{x}_t$ and $\boldsymbol{\beta}'\mathbf{x}_{t-1}$ in terms of deviations from their long-run means. Thus *deterministic* elements play an important part in the VEqCM formulation, a fact that has implications for the forecast performance of such models when there are structural breaks in deterministic terms. This issue is raised in Section 4.2.5, after a VEqCM forecast model is developed in Section 4.2.3.

4.2.3 A VEqCM Forecasting Model

In order to simplify the analysis in this chapter, and isolate the fundamental properties of the various forecasting models studied, it is helpful to assume that all parameters are known; thus estimation uncertainty is not a consider-

ation in this analysis. Further, the distance measures introduced in Section 4.1 are used to examine the properties of one-step forecasts alone, in Section 4.3, and so this section focusses on one-step forecast errors. Extending the distance measure analysis to h-step forecasts is a topic for future research.

Based on (4.5), the one-step forecast of $\Delta\mathbf{x}_{T+1}$, denoted $\Delta\widehat{\mathbf{x}}_{T+1|T}$ is given, through a well-known result, by the conditional expectation:[7]

$$\mathsf{E}_T[\Delta\mathbf{x}_{T+1}|\mathbf{x}_T] = \Delta\widehat{\mathbf{x}}_{T+1|T} = \boldsymbol{\gamma} + \boldsymbol{\alpha}(\boldsymbol{\beta}'\mathbf{x}_T - \boldsymbol{\mu}). \tag{4.7}$$

Thus the one-step forecast error for the level of $\mathbf{x}_{T+1}$ is given by:

$$\widetilde{\epsilon}_{T+1|T} = \mathbf{x}_{T+1} - \widehat{\mathbf{x}}_{T+1|T}.$$

From (4.7), it follows that:

$$\begin{aligned}\widehat{\mathbf{x}}_{T+1|T} &= \mathbf{x}_T + \boldsymbol{\gamma} + \boldsymbol{\alpha}(\boldsymbol{\beta}'\mathbf{x}_T - \boldsymbol{\mu}) \\ &= \boldsymbol{\tau} + \boldsymbol{\Gamma}\mathbf{x}_T,\end{aligned}$$

and from (4.3),

$$\mathbf{x}_{T+1} = \boldsymbol{\tau} + \boldsymbol{\Gamma}\mathbf{x}_T + \boldsymbol{\varepsilon}_{T+1},$$

and so:

$$\widetilde{\epsilon}_{T+1|T} = \boldsymbol{\varepsilon}_{T+1}.$$

Therefore $\mathsf{E}[\widetilde{\epsilon}_{T+1|T}] = \mathbf{0}$ and $\mathsf{V}[\widetilde{\epsilon}_{T+1|T}] = \boldsymbol{\Omega}_\varepsilon$.

Further, the one-step error for the growth in $\mathbf{x}_{T+1}$ is:

$$\begin{aligned}\widehat{\boldsymbol{\varepsilon}}_{T+1|T} &= \Delta\mathbf{x}_{T+1} - \Delta\widehat{\mathbf{x}}_{T+1|T} \\ &= \boldsymbol{\gamma} + \boldsymbol{\alpha}(\boldsymbol{\beta}'\mathbf{x}_T - \boldsymbol{\mu}) + \boldsymbol{\varepsilon}_{T+1} - \boldsymbol{\gamma} - \boldsymbol{\alpha}(\boldsymbol{\beta}'\mathbf{x}_T - \boldsymbol{\mu}) \\ &= \boldsymbol{\varepsilon}_{T+1},\end{aligned}$$

and so $\widehat{\boldsymbol{\epsilon}}_{T+1|T} = \widetilde{\varepsilon}_{T+1|T}$, so that for one-step forecasts, the forecast errors for levels and growth rates are equal.

In a well-specified, constant-parameter case therefore, the forecast model (4.7) will deliver unbiased forecasts with the smallest mean squared forecast

[7] See, e.g. Hamilton (1994, Chapter 4).

error matrix: the density of $\widehat{\boldsymbol{\varepsilon}}_{T+1|T}$ is in this case equivalent to the irreducible DGP innovation density $\mathsf{f}_{\varepsilon}(\boldsymbol{\varepsilon}_{T+1})$.

4.2.4 Introducing mis-specification into the DGP

For the purposes of the analysis in Section 4.3, it is interesting to allow a degree of mis-specification in the forecast model (4.7), by extending the DGP (4.3) to incorporate a larger information set. Thus the in-sample DGP could be of the form:

$$\Delta \mathbf{x}_t = \boldsymbol{\gamma}_0 + \boldsymbol{\alpha}_0(\boldsymbol{\beta}_0' \mathbf{x}_{t-1} - \boldsymbol{\mu}_0) + \boldsymbol{\Upsilon}_0 \mathbf{z}_t + \boldsymbol{\nu}_t, \tag{4.8}$$

where $\boldsymbol{\nu}_t \sim \mathsf{NID}_n(\mathbf{0}, \boldsymbol{\Omega}_{\nu})$ and following Hendry (2006), the 0 subscript denotes population parameters. Relating (4.8) to (4.3), $\boldsymbol{\nu}_t$ is the DGP innovation, and $\boldsymbol{\varepsilon}_t = \boldsymbol{\Upsilon}_0 \mathbf{z}_t + \boldsymbol{\nu}_t$.

For simplicity, it is assumed that $\mathbf{z}_t$ is an $\mathsf{I}(0)$ k-dimensional zero-mean VAR, such that:

$$\mathbf{z}_t = \boldsymbol{\Phi} \mathbf{z}_{t-1} + \boldsymbol{\eta}_t,$$

where $\boldsymbol{\eta}_t \sim \mathsf{IN}_k(\mathbf{0}, \boldsymbol{\Omega}_{\eta})$ and all the eigenvalues of $\boldsymbol{\Phi}$ lie within the unit circle; therefore $\boldsymbol{\Upsilon}_0$ is an $n \times k$ matrix of coefficients. A further, and perhaps unrealistic, simplification is that $\mathbf{z}_t$ is orthogonal to $\boldsymbol{\beta}_0' \mathbf{x}_{t-1}$, and $\boldsymbol{\beta}_0' \boldsymbol{\Upsilon}_0 = \mathbf{0}$, so that the parameter estimates in (4.8) are consistent.

4.2.5 Forecasting in the presence of structural breaks

Since an assessment of the impact of structural breaks, and the development of models robust to such changes is the central focus of this thesis, it is appropriate to introduce the propensity for some parameters in the DGP outlined in (4.8) to shift over the forecast horizon. A number of parameters could shift, including the unconditional growth rate $\boldsymbol{\gamma}$, the short-run feedback vector $\boldsymbol{\alpha}$, or the variance of the DGP innovation $\boldsymbol{\Omega}_{\nu}$. Some of these have been examined already in the wider context of forecasting and estimation (see, for example, Clements and Hendry (1998, 1999, 2006), or Kurita and Nielsen (2004)). Of interest in this section, however, are location shifts in

the equilibrium mean $\boldsymbol{\mu}$. As Chapter 3 indicated, such breaks can easily lead to systematic forecast failure in the general class of equilibrium-correction models.

Denoting the post-break equilibrium mean as $\boldsymbol{\mu}^*$, then $\nabla\boldsymbol{\mu}^* = \boldsymbol{\mu}^* - \boldsymbol{\mu}$. For a break occurring at the forecast origin, T, the DGP for $\Delta\mathbf{x}_{T+1}$ is given by:

$$\Delta\mathbf{x}_{T+1} = \boldsymbol{\gamma} + \boldsymbol{\alpha}(\boldsymbol{\beta}'\mathbf{x}_T - \boldsymbol{\mu}^*) + \boldsymbol{\Upsilon}\mathbf{z}_{T+1} + \boldsymbol{\nu}_{T+1},$$

where 0 subscripts have been omitted from coefficients for notational simplicity. Adding and subtracting $\boldsymbol{\alpha}\boldsymbol{\mu}$ yields:

$$\begin{aligned}\Delta\mathbf{x}_{T+1} &= \boldsymbol{\gamma} + \boldsymbol{\alpha}(\boldsymbol{\beta}'\mathbf{x}_T - \boldsymbol{\mu}) + \boldsymbol{\Upsilon}\mathbf{z}_{T+1} + \boldsymbol{\nu}_{T+1} - \boldsymbol{\alpha}\nabla\boldsymbol{\mu}^* \\ &= \Delta\widehat{\mathbf{x}}_{T+1|T} + \boldsymbol{\Upsilon}\mathbf{z}_{T+1} + \boldsymbol{\nu}_{T+1} - \boldsymbol{\alpha}\nabla\boldsymbol{\mu}^*.\end{aligned} \tag{4.9}$$

Since $\mathbf{z}_{T+1}$ is a vector of *omitted* variables, it does not enter into the VEqCM forecast, $\Delta\widehat{\mathbf{x}}_{T+1|T}$; however, the unconditional mean $\mathsf{E}[\mathbf{z}_{T+1}] = \mathbf{0}$, and so the expected forecast error $\widehat{\boldsymbol{\varepsilon}}_{T+1|T} = \Delta\mathbf{x}_{T+1} - \Delta\widehat{\mathbf{x}}_{T+1|T}$ is:

$$\mathsf{E}\left[\widehat{\boldsymbol{\varepsilon}}_{T+1|T}\right] = -\boldsymbol{\alpha}\nabla\boldsymbol{\mu}^*,$$

with variance:

$$\begin{aligned}\mathsf{V}\left[\widehat{\boldsymbol{\varepsilon}}_{T+1|T}\right] &= \mathsf{V}[\boldsymbol{\Upsilon}\boldsymbol{z}_{T+1} + \boldsymbol{\nu}_{T+1}] \\ &= \boldsymbol{\Upsilon}\mathsf{V}[\mathbf{z}_{T+1}]\boldsymbol{\Upsilon}' + \boldsymbol{\Omega}_\nu.\end{aligned} \tag{4.10}$$

The forecast error bias $-\boldsymbol{\alpha}\nabla\boldsymbol{\mu}^*$ reflects the unanticipated change in $\Delta\mathbf{x}_{T+1}$ resulting from the equilibrium mean change, and this bias persists for one-step forecasts as the forecast origin moves away from the break point, since $\mathsf{E}[\widehat{\boldsymbol{\varepsilon}}_{T+h|T+h-1}]$ is given by:

$$\mathsf{E}[\Delta\widehat{\mathbf{x}}_{T+h|T+h-1} + \boldsymbol{\Upsilon}\mathbf{z}_{T+h} + \boldsymbol{\nu}_{T+h} - \boldsymbol{\alpha}\nabla\boldsymbol{\mu}^* - \Delta\widehat{\mathbf{x}}_{T+h|T+h-1}] = -\boldsymbol{\alpha}\nabla\boldsymbol{\mu}^*,$$

for all $h > 0$. Therefore a location shift in $\boldsymbol{\mu}$ induces *systematic* forecast bias. Thus the need for *robust* forecast models that are not systematically affected by such structural breaks is evident: these are now developed in

Section 4.3.

4.3 Robust Forecasting Devices

Since shifts in deterministic terms are of concern here, an obvious strategy to explore is developing a robust forecast by eliminating such terms from the forecast model. A simple example can demonstrate that taking the second difference of an I(1) series will remove two unit roots, and any kind of intercept.

The variable x_t is given by a deterministic, univariate version of the simple DGP in (4.3), where $\beta = 1$, so:

$$x_t = (1 + \alpha)x_{t-1} - \alpha\mu, \tag{4.11}$$

and therefore:

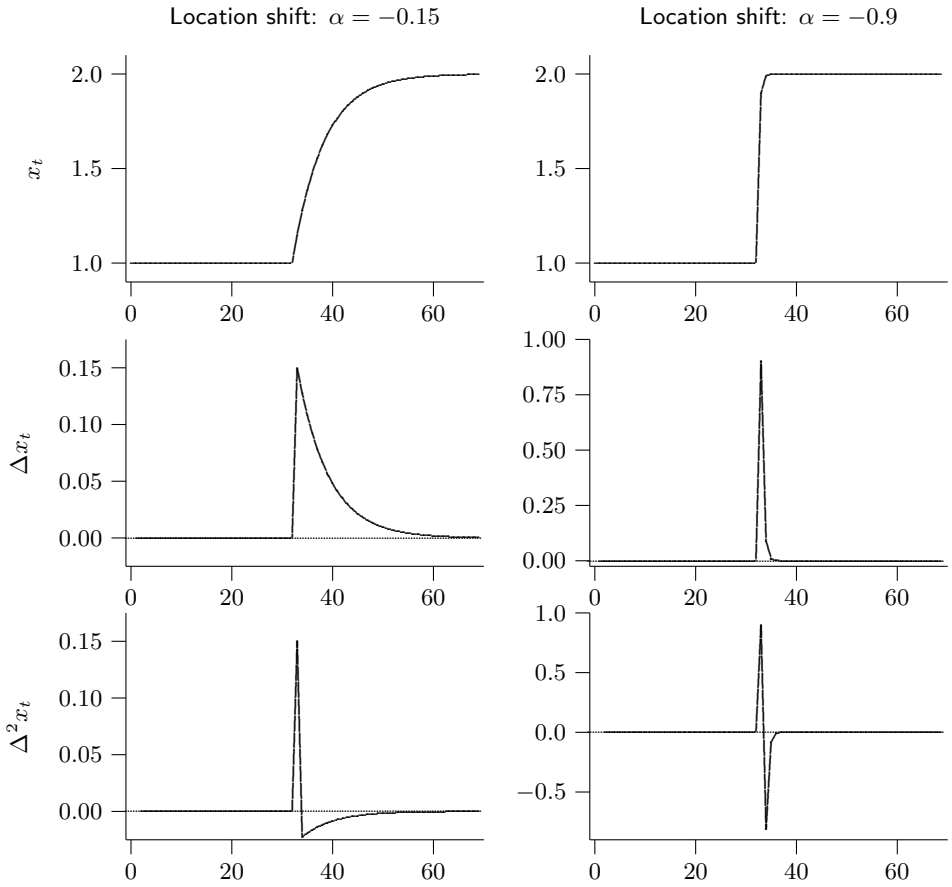
$$\Delta x_t = \alpha(x_{t-1} - \mu). \tag{4.12}$$

Figure 4.2 demonstrates two applications of this model. In both cases, the equilibrium mean μ is initially 1, before breaking at time $t = 35$ to 2: thus $\nabla\mu^* = 1$. The left hand column (A) of the figure shows the level, first difference and second difference of x_t, with a sluggish feedback parameter $\alpha = -0.15$; the right hand column (B) shows the same series, for a more responsive feedback parameter $\alpha = -0.9$.

Following a break in μ, the levels in panels A1 and B1 adjust over time towards their new equilibria. Since the speed of adjustment is lower in case A, x_t only reaches its new mean about 25 periods after the shift; in contrast, adjustment in case B occurs within three periods. The fact that the level of x_t may be affected by its pre-break history after a shift in the mean, is important when considering the various transformations below.

Row 2 in Figure 4.2 shows Δx_t, as given by (4.12). In contrast to the level of x_t, which shifted permanently after the break, the growth in x_t has an impulse at the break point, and then subsides back to its pre-break mean, at a speed determined by α. Similarly, the *second* difference $\Delta^2 x_t$ in row 3 shows that location shifts entail 'blips' of a kind, but no mean changes. The

Figure 4.2: EQUILIBRIUM LOCATION SHIFTS



NOTES: The panels refer to equation (4.11), for the level (row 1), first difference (2) and second difference (3) of the series x_t. There is a shift in the equilibrium mean of $\nabla\mu^* = 1$ at $t = 35$.

crucial difference between rows 2 and 3, however, is the fact that any forecast based on the equilibrium-correction representation in (4.12) that *does not* incorporate the new mean will be systematically biased, as Section 4.2.5 demonstrated. In this example, the one-step forecast error of the growth in x_t after the break is:

$$\begin{aligned}\mathsf{E}\left[\Delta x_{T+1} - \Delta\widehat{x}_{T+1|T}\right] &= \alpha(x_T - \mu) - \alpha\nabla\mu^* - \alpha(x_T - \mu) \\ &= -\alpha\nabla\mu^*.\end{aligned}$$

In contrast, a simple double difference-based forecast need not have the same

systematic bias, as is now discussed in Section 4.3.1.

4.3.1 Double-Differenced Devices

The underlying mechanics of a double-differenced forecast device have already been examined by Hendry (2006), and so many of the key results presented here are not new. However, it is useful to discuss the properties of a DDD, and make explicit a number of its features in relation to VEqCM and differenced-VEqCM forecast devices, that are only implicit in existing studies.

Since the DGP specified in (4.8) is I(1), it would be reasonable to assume that the series does not continuously accelerate; in this case,

$$\mathsf{E}[\Delta^2 \mathbf{x}_t] = \mathbf{0},$$

and considering row 3 of Figure 4.2, a possible forecast of $\Delta \mathbf{x}_{T+1}$ could be:

$$\Delta \widetilde{\mathbf{x}}_{T+1|T} = \Delta \mathbf{x}_T.$$

This kind of model would be more robust to deterministic shifts such as the break in μ studied above, since the unconditional forecast bias is zero.

In the absence of breaks, the forecast error $\widetilde{\boldsymbol{\varepsilon}}_{T+1|T}$ is given by:

$$\begin{aligned}
\widetilde{\boldsymbol{\varepsilon}}_{t+1|t} &= \Delta \mathbf{x}_{t+1} - \Delta \widetilde{\mathbf{x}}_{t+1|t} \\
&= \boldsymbol{\gamma} + \boldsymbol{\alpha}(\boldsymbol{\beta}' \mathbf{x}_t - \boldsymbol{\mu}) + \boldsymbol{\Upsilon} \mathbf{z}_{t+1} + \boldsymbol{\nu}_{t+1} \\
&\quad - \boldsymbol{\gamma} - \boldsymbol{\alpha}(\boldsymbol{\beta}' \mathbf{x}_{t-1} - \boldsymbol{\mu}) - \boldsymbol{\Upsilon} \mathbf{z}_t - \boldsymbol{\nu}_t \\
&= \boldsymbol{\alpha}\boldsymbol{\beta}' \Delta \mathbf{x}_t + \boldsymbol{\Upsilon} \Delta \mathbf{z}_{t+1} + \Delta \boldsymbol{\nu}_{t+1}. \qquad (4.13)
\end{aligned}$$

Therefore:

$$\mathsf{E}\left[\widetilde{\boldsymbol{\varepsilon}}_{t+1|t}\right] = \mathbf{0} \qquad (4.14a)$$

and:

$$\mathsf{V}\left[\widetilde{\boldsymbol{\varepsilon}}_{t+1|t}\right] = \boldsymbol{\alpha}\boldsymbol{\beta}' \mathsf{V}[\Delta \mathbf{x}_t] \boldsymbol{\beta}\boldsymbol{\alpha}' + \boldsymbol{\Upsilon} \mathsf{V}[\Delta \mathbf{z}_{t+1}] \boldsymbol{\Upsilon}' + 2\boldsymbol{\Omega}_\nu, \qquad (4.14b)$$

assuming covariances are zero. Compared to the variance of the VEqCM

forecast $\widehat{\boldsymbol{\varepsilon}}_{t+1|t}$ which, from equation (4.10) is:

$$\mathsf{V}\left[\widehat{\boldsymbol{\varepsilon}}_{t+1|t}\right] = \boldsymbol{\Upsilon}\mathsf{V}[\mathbf{z}_{t+1}]\boldsymbol{\Upsilon}' + \boldsymbol{\Omega}_{\nu},$$

it is not necessarily the case that (4.14b) will be greater: in the case of a highly persistent (but not I(1)) $\mathbf{z}_t$ series, the variance of the VEqCM forecast error may exceed that of the DDD.

Following a structural break in $\boldsymbol{\mu}$, however, (4.14a) is no longer correct. In the first period after the break, the DDD forecast error will be

$$\begin{aligned}
\widetilde{\boldsymbol{\varepsilon}}_{T+1|T} &= \Delta\mathbf{x}_{T+1} - \Delta\widetilde{\mathbf{x}}_{T+1|T} \\
&= \boldsymbol{\gamma} + \boldsymbol{\alpha}(\boldsymbol{\beta}'\mathbf{x}_T - \boldsymbol{\mu}) - \boldsymbol{\alpha}\nabla\boldsymbol{\mu}^* + \boldsymbol{\Upsilon}\mathbf{z}_{T+1} + \boldsymbol{\nu}_{T+1} \\
&\quad - \boldsymbol{\gamma} - \boldsymbol{\alpha}(\boldsymbol{\beta}'\mathbf{x}_T - \boldsymbol{\mu}) - \boldsymbol{\Upsilon}\mathbf{z}_T - \boldsymbol{\nu}_T \\
&= \boldsymbol{\alpha}\boldsymbol{\beta}'\Delta\mathbf{x}_T + \boldsymbol{\Upsilon}\Delta\mathbf{z}_{T+1} + \Delta\boldsymbol{\nu}_{T+1} - \boldsymbol{\alpha}\nabla\boldsymbol{\mu}^*,
\end{aligned}$$

and so

$$\mathsf{E}\left[\widetilde{\boldsymbol{\varepsilon}}_{T+1|T}\right] = -\boldsymbol{\alpha}\nabla\boldsymbol{\mu}^*.$$

This is the same as the forecast error bias in a VEqCM, and corresponds to the upward spike in the $\Delta^2 x_t$ series in row 3 of Figure 4.2. However, one period after this,

$$\begin{aligned}
\widetilde{\boldsymbol{\varepsilon}}_{T+2|T+1} &= \Delta\mathbf{x}_{T+2} - \Delta\widetilde{\mathbf{x}}_{T+2|T+1} \\
&= \boldsymbol{\gamma} + \boldsymbol{\alpha}(\boldsymbol{\beta}'\mathbf{x}_{T+1} - \boldsymbol{\mu}^*) + \boldsymbol{\Upsilon}\mathbf{z}_{T+2} + \boldsymbol{\nu}_{T+2} - \Delta\mathbf{x}_{T+1} \\
&= \boldsymbol{\gamma} + \boldsymbol{\alpha}(\boldsymbol{\beta}'\mathbf{x}_{T+1} - \boldsymbol{\mu}^*) + \boldsymbol{\Upsilon}\mathbf{z}_{T+2} + \boldsymbol{\nu}_{T+2} \\
&\quad - \boldsymbol{\gamma} - \boldsymbol{\alpha}(\boldsymbol{\beta}'\mathbf{x}_T - \boldsymbol{\mu}^*) - \boldsymbol{\nu}_{T+1} \\
&= \boldsymbol{\alpha}\boldsymbol{\beta}'\Delta\mathbf{x}_{T+1} + \boldsymbol{\Upsilon}\Delta\mathbf{z}_{T+2} + \Delta\boldsymbol{\nu}_{T+2}.
\end{aligned}$$

Since $\mathsf{E}[\Delta\mathbf{x}_{T+1}] = -\boldsymbol{\alpha}\nabla\boldsymbol{\mu}^*$, it follows that

$$\mathsf{E}\left[\widetilde{\boldsymbol{\varepsilon}}_{T+2|T+1}\right] = -\boldsymbol{\alpha}(\boldsymbol{\beta}'\boldsymbol{\alpha})\nabla\boldsymbol{\mu}^*. \tag{4.15}$$

In contrast to $\mathsf{E}\left[\widehat{\boldsymbol{\varepsilon}}_{T+2|T+1}\right] = -\boldsymbol{\alpha}\nabla\boldsymbol{\mu}^*$, (4.15) must be lower, and Appendix

4.A shows that the asymptotic bias is:

$$\mathsf{E}\left[\widetilde{\varepsilon}_{T+h|T+h-1}\right] = -\boldsymbol{\alpha}(\boldsymbol{\beta}'\boldsymbol{\alpha})\boldsymbol{\Psi}^{h-2}\nabla\boldsymbol{\mu}^* \quad \text{for } h \geq 2$$

which dies out as $h \to \infty$. This corresponds precisely to the behaviour in row 3 of Figure 4.2: after the initial impulse at the break point, there is a negative bias in the double difference forecast, which disappears shortly after the break.

That this bias exists at all once the location shift has occurred is due to the fact that even though the VEqCM is correcting towards a new equilibrium $\boldsymbol{\mu}^*$, its *level* at the time of the break will be a function of the *old* equilibrium mean: the presence of $\boldsymbol{\alpha}\boldsymbol{\beta}'\Delta\mathbf{x}_t$ in (4.13) introduces a degree of persistence in the forecast error, although this dies out quickly.

Eliminating this effect would be desirable for two reasons: first, it could attenuate the forecast bias for post-break periods, and secondly, it would reduce the *variance* of the forecast error in (4.14b); thus Section 4.3.2 now considers a differenced VEqCM forecast model.

4.3.2 Differenced VEqCM

The starting point for an alternative forecast model, as suggested by Hendry (2004, 2006) is the in-sample DGP (4.8). Taking the first difference of this yields:

$$\begin{aligned}\Delta^2\mathbf{x}_t = \Delta\mathbf{x}_t - \Delta\mathbf{x}_{t-1} &= (\boldsymbol{\gamma} - \boldsymbol{\gamma}) + (\boldsymbol{\alpha}\boldsymbol{\mu} - \boldsymbol{\alpha}\boldsymbol{\mu}) + \boldsymbol{\alpha}\boldsymbol{\beta}'(\mathbf{x}_{t-1} - \mathbf{x}_{t-2}) \\ &\quad + \boldsymbol{\Upsilon}(\mathbf{z}_t - \mathbf{z}_{t-1}) + (\boldsymbol{\nu}_t - \boldsymbol{\nu}_{t-1}) \\ &= \boldsymbol{\alpha}\boldsymbol{\beta}'\Delta\mathbf{x}_{t-1} + \boldsymbol{\Upsilon}\Delta\mathbf{z}_t + \Delta\boldsymbol{\nu}_t,\end{aligned}$$

or equivalently,

$$\Delta\mathbf{x}_t = \boldsymbol{\Gamma}\Delta\mathbf{x}_{t-1} + \boldsymbol{\Upsilon}\Delta\mathbf{z}_t + \Delta\boldsymbol{\nu}_t.$$

This extends the double-differenced VAR examined above, since $\boldsymbol{\Gamma} = \mathbf{I}_n + \boldsymbol{\alpha}\boldsymbol{\beta}'$, where the rank conditions from cointegration are now imposed.

A forecast model based on this differenced VEqCM (DVEqCM) specifi-

cation, denoted $\Delta\widetilde{\widetilde{\mathbf{x}}}_{t+1|t}$ would be given, in this simple case, by:

$$\Delta\widetilde{\widetilde{\mathbf{x}}}_{t+1|t} = (\mathbf{I}_n + \boldsymbol{\alpha}\boldsymbol{\beta}')\Delta\mathbf{x}_{t-1}, \tag{4.16}$$

leading to a forecast error $\widetilde{\widetilde{\boldsymbol{\varepsilon}}}_{t+1|t}$ of:

$$\begin{aligned}
\widetilde{\widetilde{\boldsymbol{\varepsilon}}}_{t+1|t} &= \Delta\mathbf{x}_{t+1} - \Delta\widetilde{\widetilde{\mathbf{x}}}_{t+1|t} \\
&= \Delta\mathbf{x}_{t+1} - (\mathbf{I}_n + \boldsymbol{\alpha}\boldsymbol{\beta}')\Delta\mathbf{x}_t \\
&= \Delta^2\mathbf{x}_{t+1} - \boldsymbol{\alpha}\boldsymbol{\beta}'\Delta\mathbf{x}_t \\
&= \boldsymbol{\alpha}\boldsymbol{\beta}'\Delta\mathbf{x}_t + \boldsymbol{\Upsilon}\Delta\mathbf{z}_{t+1} + \Delta\boldsymbol{\nu}_{t+1} - \boldsymbol{\alpha}\boldsymbol{\beta}'\Delta\mathbf{x}_t \\
&= \boldsymbol{\Upsilon}\Delta\mathbf{z}_{t+1} + \Delta\boldsymbol{\nu}_{t+1}.
\end{aligned}$$

Thus:

$$\mathsf{E}\left[\widetilde{\widetilde{\boldsymbol{\varepsilon}}}_{t+1|t}\right] = \mathbf{0} \tag{4.17a}$$

and:

$$\mathsf{V}\left[\widetilde{\widetilde{\boldsymbol{\varepsilon}}}_{t+1|t}\right] = \boldsymbol{\Upsilon}\mathsf{V}[\Delta\mathbf{z}_{t+1}]\boldsymbol{\Upsilon}' + 2\boldsymbol{\Omega}_\nu. \tag{4.17b}$$

Despite the presence of an extra $\boldsymbol{\Omega}_\nu$ term in (4.17b), it is not inevitable that the DVEqCM variance is greater than the VEqCM variance (4.10): in contrast to the $\mathsf{V}[\Delta\mathbf{z}_{t+1}]$ term appearing in the former, the $\mathsf{V}[\mathbf{z}_{t+1}]$ in the latter could, given appropriate assumptions regarding the persistence parameter $\boldsymbol{\Phi}$, be large enough to offset the effect of doubling $\boldsymbol{\Omega}_\nu$.

In relation to the variance of a DDD (in equation (4.14b)), an obvious difference in (4.17b) is the lack of an $\boldsymbol{\alpha}\boldsymbol{\beta}'\mathsf{V}[\Delta\mathbf{x}_t]\boldsymbol{\beta}\boldsymbol{\alpha}'$ term; thus in a cointegrated system (where $r > 0$), a DVEqCM forecast dominates a DDD in variance, although both forecast error means are equal to zero in the absence of parameter changes, for known parameters.

If there *are* breaks in parameters, then the benefits of using a DVEqCM transformation are even greater. After a change in the equilibrium mean to $\boldsymbol{\mu}^*$ from $\boldsymbol{\mu}$ at time T, the DVEqCM forecast is biased, since (using (4.9)):

$$\mathsf{E}[\Delta\mathbf{x}_{T+1}] = \boldsymbol{\gamma} - \boldsymbol{\alpha}\nabla\boldsymbol{\mu}^*,$$

and yet:

$$\mathsf{E}\left[\Delta\widetilde{\widetilde{\mathbf{x}}}_{T+1|T}\right] = \mathsf{E}\left[(\mathbf{I}_n + \boldsymbol{\alpha}\boldsymbol{\beta}')\Delta\mathbf{x}_T\right] = \boldsymbol{\gamma},$$

since $\boldsymbol{\beta}'\boldsymbol{\gamma} = \mathbf{0}$. Thus:

$$\mathsf{E}\left[\Delta\mathbf{x}_{T+1} - \Delta\widetilde{\widetilde{\mathbf{x}}}_{T+1|T}\right] = \boldsymbol{\gamma} - \boldsymbol{\alpha}\nabla\boldsymbol{\mu}^* - \boldsymbol{\gamma} = -\boldsymbol{\alpha}\nabla\boldsymbol{\mu}^*,$$

which is the same as the forecast error biases for both the VEqCM and DDD forecasts immediately after a break. Where the DVEqCM dominates these other forecast models, however, is one period after this. Since the cointegrating restrictions have been imposed, the DVEqCM forecast incorporates the new equilibrium mean at time $T+1$, as:

$$\mathbf{x}_{T+1} = \boldsymbol{\gamma} - \boldsymbol{\alpha}\boldsymbol{\mu} + (\mathbf{I}_n + \boldsymbol{\alpha}\boldsymbol{\beta}')\mathbf{x}_T + \boldsymbol{\Upsilon}\mathbf{z}_{T+1} - \boldsymbol{\alpha}\nabla\boldsymbol{\mu}^* + \boldsymbol{\nu}_{T+1},$$

and therefore:

$$\begin{aligned}
\mathsf{E}\left[\boldsymbol{\beta}'\mathbf{x}_{T+1}\right] &= \mathsf{E}[\boldsymbol{\beta}'(\mathbf{I}_n + \boldsymbol{\alpha}\boldsymbol{\beta}')\mathbf{x}_T] - \boldsymbol{\beta}'\boldsymbol{\alpha}\boldsymbol{\mu} - \boldsymbol{\beta}'\boldsymbol{\alpha}\nabla\boldsymbol{\mu}^* \\
&= \boldsymbol{\mu} + \boldsymbol{\beta}'\boldsymbol{\alpha}\boldsymbol{\mu} - \boldsymbol{\beta}'\boldsymbol{\alpha}\boldsymbol{\mu} - \boldsymbol{\beta}'\boldsymbol{\alpha}\nabla\boldsymbol{\mu}^* \\
&= \boldsymbol{\mu}^* - \nabla\boldsymbol{\mu}^* - \boldsymbol{\beta}'\boldsymbol{\alpha}\nabla\boldsymbol{\mu}^* \\
&= \boldsymbol{\mu}^* - \boldsymbol{\Psi}\nabla\boldsymbol{\mu}^*.
\end{aligned}$$

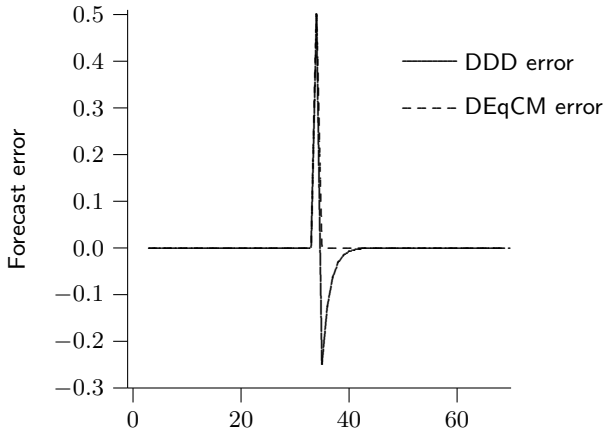
It follows that:

$$\begin{aligned}
\mathsf{E}[\Delta\mathbf{x}_{T+2}] &= \mathsf{E}[\boldsymbol{\gamma} + \boldsymbol{\alpha}(\boldsymbol{\beta}'\mathbf{x}_{T+1} - \boldsymbol{\mu}^*) + \boldsymbol{\Upsilon}\mathbf{z}_{T+2} + \boldsymbol{\nu}_{T+2}] \\
&= \boldsymbol{\gamma} - \boldsymbol{\alpha}\boldsymbol{\Psi}\nabla\boldsymbol{\mu}^*.
\end{aligned}$$

Since:

$$\begin{aligned}
\mathsf{E}\left[\Delta\widetilde{\widetilde{\mathbf{x}}}_{T+2|T+1}\right] &= \mathsf{E}[\Delta\mathbf{x}_{T+1}] + \boldsymbol{\alpha}\mathsf{E}[\boldsymbol{\beta}'\Delta\mathbf{x}_{T+1}] \\
&= \boldsymbol{\gamma} - \boldsymbol{\alpha}\nabla\boldsymbol{\mu}^* - \boldsymbol{\alpha}(\boldsymbol{\beta}'\boldsymbol{\alpha})\nabla\boldsymbol{\mu}^*,
\end{aligned}$$

Figure 4.3: LEVEL EFFECTS: DDD AND DEQCM FORECAST ERRORS COMPARED



NOTES: The figure relates to the simple univariate model introduced in Section 4.3, with feedback parameter $\alpha = -0.5$, with the forecast errors of a double-differenced device (DDD) compared to those of a differenced EqCM transformation (DEqCM).

the expectation of the forecast error $\widetilde{\widetilde{\boldsymbol{\varepsilon}}}_{T+2|T+1}$ is given by

$$\begin{aligned}\mathsf{E}\left[\widetilde{\widetilde{\boldsymbol{\varepsilon}}}_{T+2|T+1}\right] &= \mathsf{E}\left[\Delta\mathbf{x}_{T+2} - \Delta\widetilde{\widetilde{\mathbf{x}}}_{T+2|T+1}\right] \\ &= \boldsymbol{\gamma} - \boldsymbol{\alpha}\boldsymbol{\Psi}\nabla\boldsymbol{\mu}^* - \boldsymbol{\gamma} + \boldsymbol{\alpha}\nabla\boldsymbol{\mu}^* + \boldsymbol{\alpha}(\boldsymbol{\beta}'\boldsymbol{\alpha})\nabla\boldsymbol{\mu}^* \\ &= (\boldsymbol{\alpha} - \boldsymbol{\alpha} - \boldsymbol{\alpha}(\boldsymbol{\beta}'\boldsymbol{\alpha}) + \boldsymbol{\alpha}(\boldsymbol{\beta}'\boldsymbol{\alpha}))\nabla\boldsymbol{\mu}^* = \mathbf{0}.\end{aligned}$$

Thus the bias in the DDD that dies out only asymptotically, disappears *immediately* after the break in the DVEqCM forecast: the absence of any persistent level effects ensures that not only is the variance lower (as equation (4.17b) demonstrates), but also forecast bias dies out immediately. This difference is illustrated in Figure 4.3, which shows the DDD and DEqCM forecast errors from the simple univariate model discussed at the beginning of Section 4.3. Both forecast models suffer the same error at the time of the break, equal to $-\alpha\nabla\mu^*$; only one period later, the DEqCM forecast error returns to zero, whilst the DDD forecast error undershoots, before returning to zero relatively quickly.

4.3.3 Robust Devices and Error Metrics

Three forecast models have been discussed so far in this chapter: a conventional VEqCM-based model, a double-differenced device and a differenced VEqCM transformation. Following the comparison of the first two moments of each model's forecast errors, an obvious direction for this section is to use the measures of affinity developed in Section 4.1 to evaluate the forecast error densities of the three models established above.

To this end, a simple scalar model similar to equation (4.12) is helpful, to implement the density error metric comparison. When $n = k = 1$, the DGP is given by:

$$\Delta x_t = \gamma + \alpha(x_{t-1} - \mu) + \lambda z_t + \nu_t, \tag{4.18a}$$

where $\nu_t \sim \mathsf{NID}\left(0, \sigma_\nu^2\right)$ and:

$$z_t = \phi z_{t-1} + \eta_t, \qquad \begin{array}{l} \eta_t \sim \mathsf{NID}\left(0, \sigma_\eta^2\right) \\ |\phi| < 1 \end{array} \tag{4.18b}$$

and so $\mathsf{E}[z_t] = 0$ and $\mathsf{V}[z_t] = \frac{\sigma_\eta^2}{1-\phi^2}$, with $z_0 \sim \mathsf{N}\left(0, \frac{\sigma_\eta^2}{1-\phi^2}\right)$.

In this case, the forecast models, and their respective error means and variances are given below. For the sake of clarity, the derivation of forecast error variances for the DDD and DEqCM models is presented in Appendix 4.B.

EqCM forecasts

$$\Delta \widehat{x}_{t+1|t} = \gamma + \alpha(x_t - \mu), \tag{4.19a}$$

corresponding to equation (4.7). Further,

$$\mathsf{E}\left[\widehat{\varepsilon}_{t+1|t}\right] = \mathsf{E}\left[\Delta x_{t+1} - \Delta \widehat{x}_{t+1|t}\right] = 0,$$

and

$$\mathsf{V}\left[\widehat{\varepsilon}_{t+1|t}\right] = \lambda^2 \mathsf{V}[z_t] + \sigma_\nu^2 = \lambda^2 \frac{\sigma_\eta^2}{1-\phi^2} + \sigma_\nu^2$$

DDD FORECASTS

$$\Delta\widetilde{x}_{t+1|t} = \Delta x_t, \tag{4.19b}$$

corresponding to equation (4.13). Further,

$$\mathsf{E}\left[\widetilde{\varepsilon}_{t+1|t}\right] = \mathsf{E}\left[\Delta x_{t+1} - \Delta\widetilde{x}_{t+1|t}\right] = 0,$$

and

$$\begin{aligned}\mathsf{V}\left[\widetilde{\varepsilon}_{t+1|t}\right] &= \alpha^2\mathsf{V}[\Delta x_t] + \lambda^2\mathsf{V}[\Delta z_{t+1}] + 2\sigma_\nu^2 \\ &= 2\alpha^2\left(\frac{\lambda^2\sigma_\eta^2(1-\phi^2)^{-1} + \sigma_\nu^2}{2+\alpha}\right) + 2\lambda^2\frac{\sigma_\eta^2}{1+\phi} + 2\sigma_\nu^2.\end{aligned}$$

DEqCM FORECASTS

$$\Delta\widetilde{\widetilde{x}}_{t+1|t} = (1+\alpha)\Delta x_t, \tag{4.19c}$$

corresponding to equation (4.16). Further,

$$\mathsf{E}\left[\widetilde{\widetilde{\varepsilon}}_{t+1|t}\right] = \mathsf{E}\left[\Delta x_{t+1} - \Delta\widetilde{\widetilde{x}}_{t+1|t}\right] = 0,$$

and

$$\mathsf{V}\left[\widetilde{\widetilde{\varepsilon}}_{t+1|t}\right] = \lambda^2\mathsf{V}[\Delta z_{t+1}] + 2\sigma_\nu^2 = 2\lambda^2\frac{\sigma_\eta^2}{1+\phi} + 2\sigma_\nu^2.$$

Given the assumption of normality in all relevant error terms, it follows for this simple model and the forecast devices, that the conditional and unconditional distributions of the forecast errors are all normally distributed. Thus with $\nu_t \sim \mathsf{NID}\left(0, \sigma_\nu^2\right)$ representing the baseline DGP innovation error, the forecast models (4.19a), (4.19b) and (4.19c) have densities that can be evaluated using the distance measures developed above.

The purpose of this exercise is not to prove a general result demonstrating the unequivocal benefits of using a device or transformation over a simple VEqCM forecast model. Even with a simple scalar model, there are too many free parameters that need to be fixed arbitrarily to make any kind of result 'valuable' from the point of view of selecting a particular forecast model. Rather, the intention is to underline the fact that the choice of fore-

cast model is very much a *strategic* one. In the absence of mis-specification, using a robust device entails a 'forecast cost' in terms of a greater forecast error variance (fatter tails in the forecast error density); however, as Section 4.3.3.2 below suggests, this insurance cost may be justified by the improved robustness to breaks, in the presence of even relatively small structural shifts and model mis-specification. A forecast strategy therefore involves a balance of the cost of unnecessary insurance, against the substantial benefit of being insured when there are breaks.

There are two types of situation that are of interest in this case. In the first, the effect of model mis-specification is examined, by varying the properties of the z_t process in (4.18a). Secondly, the impact of structural breaks on forecast error densities is assessed, for cases of well-specified and mis-specified models.

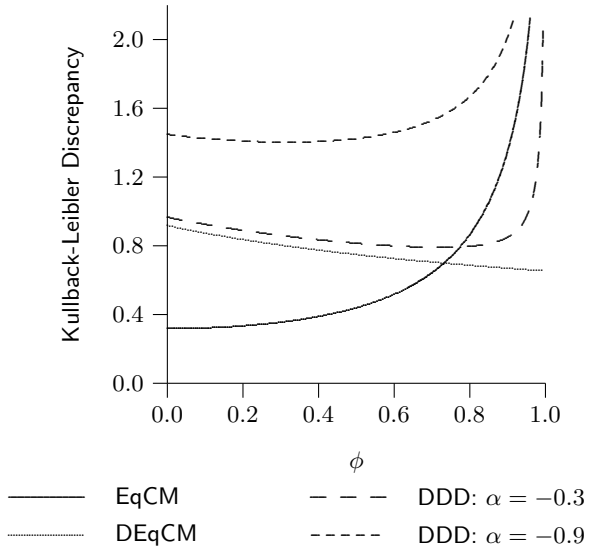
4.3.3.1 The effect of mis-specification

Allowing for mis-specification implies that $\lambda \neq 0$ in equation (4.18a). The *nature* of mis-specification can be altered by changing the persistence parameter ϕ in (4.18b) from one limit of $\phi = 0$, where z_t is white noise, up to $|\phi| = 1$, where z_t is a random walk. Since it is assumed that $z_t \sim \mathsf{I}(0)$, and in order to avoid cycles in z_t, the range of values considered for ϕ is restricted, such that $\phi \in [0, 1)$.

This is important, since the nature of mis-specification can have implications for the behaviour of conditional forecast errors. In a situation where ϕ is close to 1, the error in a simple VEqCM model will be highly auto-correlated, with a non-zero conditional expectation, even if the unconditional expectation is still zero.

To explore the effect of varying ϕ using a discrepancy measure, some assumptions need to be made regarding the variances of the error terms ν_t (the DGP innovation) and η_t (the z_t innovation), the parameter λ and the feedback parameter α. For simplicity, it is assumed that $\sigma_\nu^2 = \sigma_\eta^2 = 1$, and $\lambda = 1$. Since α affects the variance of the DDD alone, the most efficient way of demonstrating its effect is to include two curves in Figure 4.4, one where

Figure 4.4: EFFECTS OF MIS-SPECIFICATION



NOTES: The curves show the Kullback-Leibler relative entropy in a constant parameter case, as the autoregressive coefficient ϕ determining the persistence in the omitted variable z_t varies from $\phi = 0$ (the case of 'white noise misspecification') towards $\phi = 1$ (the case of 'unit root misspecification'). *Mathematica code:* `GenDistance.nb`

$\alpha = -0.3$ (sluggish feedback), and one where $\alpha = -0.9$ (rapid adjustment).[8]

The results in Figure 4.4 highlight a number of important properties of both the conventional EqCM forecasts, and various robust devices.

First, for the EqCM forecast, I_{KL} is monotonically increasing in ϕ. This is unsurprising, as the forecast variance includes a term in $(1-\phi^2)^{-1}$, which becomes a magnification factor for all $\phi > 0$, increasing from 1 up to infinity as ϕ moves from 0 to 1.

Secondly, there *are* plausible values for ϕ and α for which a simple double-differenced device has a smaller discrepancy than a conventional EqCM forecast. This relates back to the discussion on page 95, which noted that it could be possible for the variance of a VEqCM model to exceed that

[8] Johansen (1995, p. 46) observes that in a univariate model, $-2 < \alpha < 0$ is a condition for x_t being I(1) rather than I(2).

of a DDD; it is this result that underlines the behaviour observed in Figure 4.4. Of course, the DDD discrepancy also tends towards infinity as ϕ tends to 1, and with a sluggish feedback parameter, it could be the case that the EqCM forecast always dominates a DDD forecast in variance; but there *could* be a situation in which a DDD would dominate a VEqCM as a forecasting model, even without structural breaks.

Thirdly, a DEqCM forecast appears to behave in a notably different way to the previous two models. As Figure 4.4 shows, the discrepancy is monotonically decreasing in ϕ, as the magnification factor of the σ_η^2 element in the forecast error variance falls from 2 ($\phi = 0$) to 1 ($\phi = 1$). This suggests that there *must* exist a value of ϕ for which a DEqCM transformation dominates an EqCM forecast in variance; this is demonstrated in Proposition 4.3.1 below.

Proposition 4.3.1 *With a data generating process given by* (4.18a) *and* (4.18b)*, and EqCM and DEqCM forecasts given by* (4.19a) *and* (4.19c) *respectively, and providing:*

(i) $0 < \sigma_\nu^2 < \infty$, $0 < \sigma_\eta^2 < \infty$ *and* $\lambda \neq 0$, *and;*

(ii) $\nu_t \sim \mathsf{NID}\left(0, \sigma_\nu^2\right)$ *and* $\eta_t \sim \mathsf{NID}\left(0, \sigma_\eta^2\right)$

there exists $\phi \in [0, 1)$ *such that*

$$\mathsf{V}\left[\widetilde{\widetilde{\varepsilon}}_{t+1|T}\right] < \mathsf{V}\left[\widehat{\varepsilon}_{t+1|T}\right].$$

Proof. A contradiction of the proposition implies that

$$2\lambda^2 \frac{\sigma_\eta^2}{1+\phi} + 2\sigma_\nu^2 > \lambda^2 \frac{\sigma_\eta^2}{1-\phi^2} + \sigma_\nu^2$$

for $\phi \in [0, 1)$. This implies that

$$\frac{\sigma_\nu^2}{\sigma_\eta^2} > \lambda^2 \left(\frac{1}{1-\phi^2} - \frac{2}{1+\phi}\right),$$

and so

$$\frac{\sigma_\nu^2}{\sigma_\eta^2} > \lambda^2 \left(\frac{2\phi - 1}{1-\phi^2}\right).$$

Since

$$\lim_{\phi \to 1} \left(\frac{2\phi - 1}{1 - \phi^2} \right) = \infty,$$

the inequality only holds in the limit if $\sigma_\nu^2 = \infty$ or $\sigma_\eta^2 = 0$, or $\lambda = 0$, which violates condition (i); thus the Proposition must hold. ■

The fact that the mis-specification error component of the DEqCM variance becomes approximately white noise as ϕ tends to 1 therefore implies that there *must* exist a value of ϕ for which a DEqCM forecast dominates an EqCM forecast, although for high values of $\frac{\sigma_\nu^2}{\sigma_\eta^2}$, this could entail some serious mis-specification in the original EqCM.

Relating this to the discrepancy measures displayed in Figure 4.4, Lemma 4.3.1 follows on from Proposition 4.3.1:

Lemma 4.3.1 *Assuming the conditions in Proposition 4.3.1 hold, for the two unbiased forecast errors* $\widehat{\varepsilon}_{t+1|t}$ *and* $\widetilde{\widetilde{\varepsilon}}_{t+1|t}$*, there exists* $\phi \in [0, 1)$ *such that*

$$I_{KL}\big(\nu_t, \widehat{\varepsilon}_{t+1|t}\big) > I_{KL}\big(\nu_t, \widetilde{\widetilde{\varepsilon}}_{t+1|t}\big),$$

where I_{KL} *is the Kullback-Leibler discrepancy, and* ν_t *is defined in Proposition 4.3.1.*

Proof. This result follows immediately from the fact that, for a true density $f \sim \mathsf{N}\left(\mu, \sigma^2\right)$ and two alternative densities $g_1 \sim \mathsf{N}\left(\mu, \sigma_1^2\right)$ and $g_2 \sim \mathsf{N}\left(\mu, \sigma_2^2\right)$ $\forall \mu$, where $\sigma^2 \leq \sigma_1^2 < \sigma_2^2$, then $I_{KL}(f, g_1) < I_{KL}(f, g_2)$. ■

Therefore a crossing point of the error discrepancies of the EqCM and DEqCM forecasts must exist, for some value of ϕ, suggesting that for a sufficiently mis-specified model, it is better to use this transformation rather than the underlying 'structural' model. The intuition behind this is clear: the process of differencing a VEqCM model renders the mis-specification component $\mathsf{I}(-1)$ if $\phi = 0$, and it only becomes $\mathsf{I}(0)$ if $\phi = 1$; in contrast, z_t is already $\mathsf{I}(0)$ in (4.8) and in the limit as $\phi \to 1$, it becomes $\mathsf{I}(1)$. Thus in the former, the 'mis-specification noise' is decreasing in ϕ but in the latter, it is increasing.

4.3.3.2 Structural Breaks

Extending the model to allow for location shifts in μ provides some further support for the use of robust forecasting devices. The DGP outlined in (4.18a) changes, after a break in the equilibrium mean to μ^* at time T, to:

$$\Delta x_{T+1} = \gamma + \alpha(x_T - \mu) + \lambda z_{T+1} - \alpha\nabla\mu^* + \nu_{T+1},$$

and so the one-step forecast error at period T for each model has expectation:

$$\mathsf{E}\left[\widehat{\varepsilon}_{T+1|T}\right] = \mathsf{E}\left[\widetilde{\varepsilon}_{T+1|T}\right] = \mathsf{E}\left[\widetilde{\widetilde{\varepsilon}}_{T+1|T}\right] = -\alpha\nabla\mu^*.$$

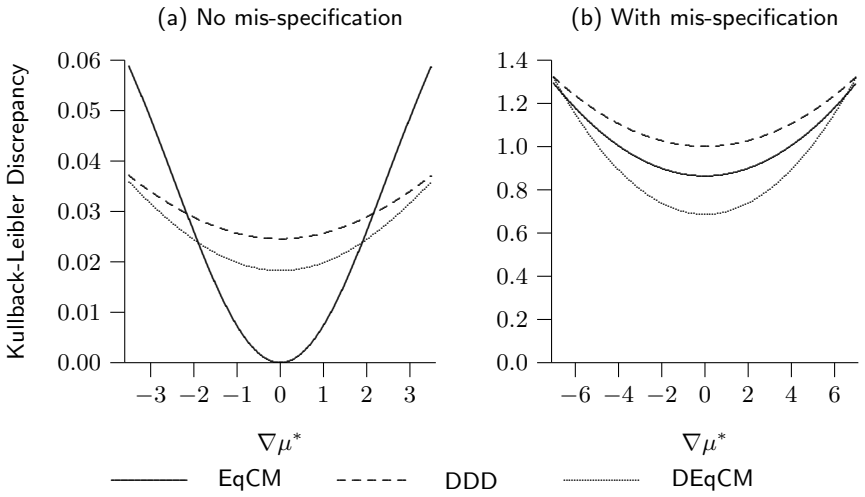
The variance of each model remains unchanged from Section 4.3.3.1, and so the mean of each model alone, changes with $\nabla\mu^*$. Continuing with the specification in 4.3.3.1, it is assumed that $\sigma_\nu^2 = \sigma_\eta^2 = 1$, and the pre-break mean is set to $\mu = 0$, so that $\nabla\mu^* = \mu^*$, and the absolute value of a shift in μ is also the relative value, when scaled by the standard deviation of the error term ν_t. Further, $\alpha = -0.5$, although a range of values is considered for an EqCM forecast below.

Based on these assumptions, Figure 4.5 shows the Kullback-Leibler discrepancy of the three forecast models' error densities, relative to the DGP innovation density of ν_t. In the left hand panel (a), there is no model mis-specification, and so $\lambda = 0$; in the right hand panel (b), there is substantial mis-specification, given by $\lambda = 1$ and $\phi = 0.9$.

Addressing panel (a), there a number of significant features to report. For all three models, any location shift in μ increases the Kullback-Leibler discrepancy, as might be expected: for a comparison of (symmetric) normal distributions, the distance between two density functions is minimised when they share the same mode, taking the variance of each as given. The difference in the discrepancy when $\nabla\mu^* = 0$ therefore reflects differences in the error variances alone, as Lemma 4.3.1 implies: in the absence of mis-specification, $\mathsf{V}\left[\widehat{\varepsilon}_{T+1|T}\right] < \mathsf{V}\left[\widetilde{\widetilde{\varepsilon}}_{T+1|T}\right] < \mathsf{V}\left[\widetilde{\varepsilon}_{T+1|T}\right]$.

However, the *rate* at which the discrepancy increases as the size of the location shift increases varies across forecast models: the EqCM error in-

Figure 4.5: A COMPARISON OF FORECAST MODELS



NOTES: Both panels show the Kullback-Leibler discrepancies for one-step forecast errors of three forecast models, at the point of a break in the equilibrium mean. The size of the break is given by $\nabla\mu^*$, which is normalised so that the (absolute) value of $\nabla\mu^*$ is also the relative break with regard to the standard deviation of the DGP innovation error ν_t. Thus $\nabla\mu^* = 1$, for instance, represents a one standard deviation shift in the mean. *Mathematica code reference:* `Gendistance.nb`

creases most rapidly, then the DEqCM error, and finally the DDD error. As a result, the discrepancy measures actually cross each other for large enough changes in μ. Thus for example, a break in the mean of just under two standard deviations is required for the DEqCM discrepancy to be lower than the EqCM distance.

The reason for this relates back to the discussion in Section 4.1.1.1. Since the variances of both the DDD and DEqCM models exceed that of the EqCM forecast, there will be a point at which the location shift $\nabla\mu^*$ is large enough to guarantee that the error discrepancy arising from a device is lower than from the conventional EqCM model. This is a powerful result, since it suggests that even without mis-specification, the use of a transformation may be justified if there are structural breaks: in panel (a), a break of only two standard deviations is required.

Adding mis-specification in panel (b) introduces the effects discussed in Section 4.3.3.1. For $\lambda = 1$ and $\phi = 0.9$, with $\alpha = -0.5$, the DEqCM discrepancy is lower than the EqCM distance, which is in turn below that

of the DDD, for location shifts of up to seven standard deviations. This appears to cement the importance of transformations in producing forecasts that minimise vulnerability to structural breaks

However, the strength of forecast devices is underlined if *robustness* is considered. At the beginning of the chapter, a robust forecast was introduced as one whose cost or discrepancy does not increase *permanently* after a structural break. Section 4.3.2 demonstrated that one period after a break, the bias in a VEqCM forecast disappears, and Section 4.3.1 demonstrated that forecast bias disappears asymptotically in a DDD model; thus shortly after the location shift[9] the discrepancy arising from either of these devices falls back to the level at $\nabla\mu^* = 0$ in Figure 4.5. In contrast, the bias in a VEqCM forecast is permanent; the increase in the discrepancy of the EqCM error in Figure 4.5 does not disappear, even asymptotically. Within the class of equilibrium-correction models (i.e. when $\alpha < 0$), a cost is incurred only if there is a location shift; this underlines the analysis in Clements and Hendry (1998, 1999, 2006) which highlights the importance of breaks in *deterministic* elements in explaining forecast failure.

Therefore the success of the DDD and DVEqCM devices is also highlighted here: both are robust to this kind of location shift, since the discrepancy in forecast error densities *does not* increase permanently after a break. Even if VEqCM forecasts have a smaller relative entropy in the absence of breaks and mis-specification, the benefits of a robust device if there *are* breaks are substantial.

4.4 Concluding comments

A central issue in this chapter was the extent to which the forecast error densities of various models compared to the density of the baseline, irreducible, innovation in a data generating process. The focus was on the class of equilibrium-correction models with an integrated-cointegrated structure.

In order to examine this, discrepancy or distance metrics were introduced, which quantified the global distance between two densities. Using

[9] Or immediately after, in the case of a VEqCM forecast.

the Kullback-Leibler information criterion, a robust model on this measure does not entail a permanent increase in its discrepancy after a break. In this light, a number of differences emerge among models.

Generally, transformations, such as a double-differenced device, or a differenced VEqCM model, are robust to breaks; although each suffers from bias at the point of a shift, this simply cannot be avoided in the absence of prior warning. Further, bias disappears either quickly, or immediately, depending on the transformation. In contrast, a conventional VEqCM model entails *permanent* bias after a break. The cost of this robustness lies in the forecast error variance, which in the absence of mis-specification, is higher for any device than the VEqCM model. However, introducing a sufficient degree of mis-specification can change this result, to the extent that given appropriate conditions, a DVEqCM transformation actually has a lower variance than a conventional VEqCM forecast mechanism.

The implications of this suggest that the choice of forecast model is very much a strategic one, since it involves balancing the cost of insurance against the cost of breaks. In the examples studied, even small breaks were sufficient to make a device worthwhile, with a DVEqCM model performing particularly well.

4.A Appendix: Forecast bias

Consider first the level of $\mathbf{x}_{T+j}$, noting that $\boldsymbol{\Gamma} = (\mathbf{I}_p + \boldsymbol{\alpha}\boldsymbol{\beta}')$ and $\boldsymbol{\Gamma}\boldsymbol{\gamma} = \boldsymbol{\gamma}$:

$$\begin{aligned}
\mathbf{x}_{T+j} &= \boldsymbol{\gamma} + \boldsymbol{\Gamma}\mathbf{x}_{T+j-1} - \boldsymbol{\alpha}\boldsymbol{\mu}^* + \boldsymbol{\epsilon}_{T+j} \\
&= \boldsymbol{\gamma} - \boldsymbol{\alpha}\boldsymbol{\mu}^* + \boldsymbol{\Gamma}\left[\boldsymbol{\gamma} + \boldsymbol{\Gamma}\mathbf{x}_{T+j-2} - \boldsymbol{\alpha}\boldsymbol{\mu}^* + \boldsymbol{\epsilon}_{T+j-1}\right] + \boldsymbol{\epsilon}_{T+j} \\
&\vdots \\
&= j\boldsymbol{\gamma} + \boldsymbol{\Gamma}^j\mathbf{x}_T - \boldsymbol{\Gamma}^{j-1}\boldsymbol{\alpha}\boldsymbol{\mu}^* + \sum_{i=0}^{j-1}\boldsymbol{\Gamma}^i\boldsymbol{\epsilon}_{T+j-i}.
\end{aligned}$$

Since $\boldsymbol{\tau} = \boldsymbol{\gamma} - \boldsymbol{\alpha}\boldsymbol{\mu}^*$, this can be written as:

$$\mathbf{x}_{T+j} = \boldsymbol{\gamma} - \boldsymbol{\alpha}\boldsymbol{\mu}^* + \boldsymbol{\Gamma}^j\mathbf{x}_T - \boldsymbol{\Gamma}^{j-1}\boldsymbol{\alpha}\boldsymbol{\mu}^* + \sum_{i=0}^{j-1}\boldsymbol{\Gamma}^i\boldsymbol{\epsilon}_{T+j-i}$$

It can be shown that:

$$\boldsymbol{\Gamma}^k = (\mathbf{I}_p - \boldsymbol{\alpha}(\boldsymbol{\beta}'\boldsymbol{\alpha})^{-1}\boldsymbol{\beta}') + \boldsymbol{\alpha}(\boldsymbol{\beta}'\boldsymbol{\alpha})^{-1}\boldsymbol{\Psi}^k\boldsymbol{\beta}',$$

where $\boldsymbol{\Psi} = (\mathbf{I}_r + \boldsymbol{\beta}'\boldsymbol{\alpha})$ corresponds to the elements of the cointegrated system whose eigenvalues lie within the unit circle (such that $\lim_{k\to\infty}\boldsymbol{\Psi}^k = \mathbf{0}$). This implies that $\boldsymbol{\Gamma}^{j-1}\boldsymbol{\alpha}\boldsymbol{\mu}^*$ can be written as:

$$\begin{aligned}
\boldsymbol{\Gamma}^{j-1}\boldsymbol{\alpha}\boldsymbol{\mu}^* &= \left(\mathbf{I}_p - \boldsymbol{\alpha}(\boldsymbol{\beta}'\boldsymbol{\alpha})^{-1}\boldsymbol{\beta}' + \boldsymbol{\alpha}(\boldsymbol{\beta}'\boldsymbol{\alpha})^{-1}\boldsymbol{\Psi}^{j-1}\boldsymbol{\beta}'\right)\boldsymbol{\alpha}\boldsymbol{\mu}^* \\
&= \boldsymbol{\alpha}(\boldsymbol{\beta}'\boldsymbol{\alpha})^{-1}\boldsymbol{\Psi}^{j-1}\boldsymbol{\beta}'\boldsymbol{\alpha}\boldsymbol{\mu}^* \\
&= \boldsymbol{\alpha}(\boldsymbol{\beta}'\boldsymbol{\alpha})^{-1}(\boldsymbol{\beta}'\boldsymbol{\alpha})\boldsymbol{\Psi}^{j-1}\boldsymbol{\mu}^* \\
&= \boldsymbol{\alpha}\boldsymbol{\Psi}^{j-1}\boldsymbol{\mu}^*.
\end{aligned}$$

Using the error-correction form:

$$\Delta\mathbf{x}_{T+j} = \boldsymbol{\gamma} + \boldsymbol{\alpha}(\boldsymbol{\beta}'\mathbf{x}_{T+j-1} - \boldsymbol{\mu}^*) + \epsilon_{T+j},$$

the resulting DDD forecast is $\widetilde{\Delta\mathbf{x}}_{T+j|T+j-1} = \Delta\mathbf{x}_{T+j-1}$, which has a one-step forecast error $\widetilde{\boldsymbol{\varepsilon}}_{T+j|T+j-1}$ given by:

$$\begin{aligned}
\widetilde{\boldsymbol{\varepsilon}}_{T+j|T+j-1} &= \Delta\mathbf{x}_{T+j} - \widetilde{\Delta\mathbf{x}}_{T+j|T+j-1} \\
&= \boldsymbol{\alpha}\boldsymbol{\beta}'\Delta\mathbf{x}_{T+j-1} + \Delta\boldsymbol{\epsilon}_{T+j} \\
&= \boldsymbol{\alpha}\boldsymbol{\beta}'\left(\boldsymbol{\gamma} + \boldsymbol{\alpha}(\boldsymbol{\beta}'\mathbf{x}_{T+j-2} - \boldsymbol{\mu}^*) + \epsilon_{T+i-1}\right) + \Delta\boldsymbol{\epsilon}_{T+j} \\
&= \boldsymbol{\alpha}\boldsymbol{\beta}'\left(\boldsymbol{\gamma} - \boldsymbol{\alpha}\boldsymbol{\mu}^* - \boldsymbol{\alpha}\boldsymbol{\beta}'\boldsymbol{\alpha}(\mathbf{I}_r + \boldsymbol{\Psi}^{j-3})\boldsymbol{\mu}^*\right) \\
&\quad + \boldsymbol{\alpha}(\boldsymbol{\beta}'\boldsymbol{\alpha})\boldsymbol{\beta}'\left[\mathbf{I}_p - \boldsymbol{\alpha}(\boldsymbol{\beta}'\boldsymbol{\alpha})^{-1}\boldsymbol{\beta}' + \boldsymbol{\alpha}(\boldsymbol{\beta}'\boldsymbol{\alpha})\boldsymbol{\Psi}^{j-2}\boldsymbol{\beta}'\right]\mathbf{x}_T \\
&\quad + \boldsymbol{\nu}_{T+j},
\end{aligned}$$

where $\boldsymbol{\nu}_{T+j}$ is a composite error term with zero mean. Thus:

$$\widetilde{\boldsymbol{\varepsilon}}_{T+j|T+j-1} = \boldsymbol{\alpha}(\boldsymbol{\beta}'\boldsymbol{\alpha})\boldsymbol{\Psi}^{j-2}(\boldsymbol{\beta}'\mathbf{x}_T) - \boldsymbol{\alpha}(\boldsymbol{\beta}'\boldsymbol{\alpha})\boldsymbol{\Psi}^{j-2}\boldsymbol{\mu}^* + \boldsymbol{\nu}_{T+j}.$$

Taking expectations, recalling that at time T, $\mathsf{E}[\boldsymbol{\beta}'\mathbf{x}_T] = \boldsymbol{\mu}$, yields:

$$\begin{aligned}\mathsf{E}\left[\widetilde{\boldsymbol{\varepsilon}}_{T+j|T+j-1}\right] &= \boldsymbol{\alpha}(\boldsymbol{\beta}'\boldsymbol{\alpha})\boldsymbol{\Psi}^{j-2}\boldsymbol{\mu} - \boldsymbol{\alpha}(\boldsymbol{\beta}'\boldsymbol{\alpha})\boldsymbol{\Psi}^{j-2}\boldsymbol{\mu}^* \\ &= -\boldsymbol{\alpha}(\boldsymbol{\beta}'\boldsymbol{\alpha})\boldsymbol{\Psi}^{j-2}\nabla\boldsymbol{\mu}^*,\end{aligned}$$

as written in the text.

4.B Appendix: Derivation of forecast error variances

4.B.1 DEqCM error variance

Since

$$\Delta z_{t+1} = (\phi - 1)z_t + \eta_{t+1},$$

it follows that

$$\begin{aligned}\mathsf{V}[\Delta z_{t+1}] &= (\phi-1)^2\frac{\sigma_\eta^2}{1-\phi^2} + \sigma_\eta^2 \\ &= \frac{-(\phi-1)\sigma_\eta^2 + (1+\phi)\sigma_\eta^2}{1+\phi} \\ &= \frac{2\sigma_\eta^2}{1+\phi}.\end{aligned}$$

Thus

$$\begin{aligned}\mathsf{V}\left[\widetilde{\widetilde{\varepsilon}}_{t+1|t}\right] &= \lambda^2\mathsf{V}[\Delta z_{t+1}] + 2\sigma_\nu^2 \\ &= 2\lambda^2\frac{\sigma_\eta^2}{1+\phi} + 2\sigma_\nu^2,\end{aligned}$$

as in the text.

4.B.2 DDD error variance

From (4.18a), and assuming (without affecting the underlying result) that $\gamma = 0$, let:

$$\begin{aligned}x_t &= (1+\alpha)x_{t-1} + \lambda z_t + \nu_t \\ &= \sum_{i=0}^{\infty}(1+\alpha)^i\left\{\lambda z_{t-i} + \nu_{t-i}\right\}.\end{aligned}$$

Then

$$\mathsf{V}[x_t] = \mathsf{E}\left[\sum_{i=0}^{\infty}(1+\alpha)^{2i}\left\{\lambda^2 z_{t-i}^2 + \nu_{t-i}^2\right\}\right],$$

by the assumption that $z \perp\!\!\!\perp \nu$. Therefore

$$\mathsf{V}[x_t] = \frac{1}{1-(1+\alpha)^2}\left(\sigma_\nu^2 + \lambda^2\sigma_\eta^2(1-\phi^2)^{-1}\right),$$

and

$$\begin{aligned}\mathsf{V}[\Delta x_t] &= \alpha^2\mathsf{V}[x_{t-1}] + \lambda^2\mathsf{V}[z_t] + \sigma_\nu^2 \\ &= -\frac{\alpha\left(\sigma_\nu^2 + \lambda^2\sigma_\eta^2(1-\phi^2)^{-1}\right)}{2+\alpha} + \lambda^2\sigma_\eta^2(1-\phi^2)^{-1} + \sigma_\nu^2 \\ &= 2\left(\frac{\lambda^2\sigma_\eta^2(1-\phi^2)^{-1} + \sigma_\nu^2}{2+\alpha}\right).\end{aligned}$$

Collecting results:

$$\mathsf{V}\left[\widetilde{\varepsilon}_{t+1|t}\right] = 2\alpha^2\left(\frac{\lambda^2\sigma_\eta^2(1-\phi^2)^{-1} + \sigma_\nu^2}{2+\alpha}\right) + 2\lambda^2\frac{\sigma_\eta^2}{1+\phi} + 2\sigma_\nu^2,$$

as in the main text.

Chapter 5
Forecast-Error Correction

> Two tourists, driving through the English countryside in search of a local village, stopped to ask a passer-by for directions. 'Ah,' was the reply, 'if I were you, I wouldn't start from here.'
>
> ANON.

When a forecast model fails, and delivers a large error, early diagnosis of the cause is essential. Whether the failure was generated by a structural shift at the forecast origin, or by an inaccurate observation over the forecast horizon, has implications for the most appropriate response: since if there *has* been a structural break, the model should correct for this rapidly, but if there is purely measurement error, then forecast failure can be discounted.

Chapter 4 showed that in a straightforward world free of measurement error, the size of a forecast error when there is a break conveys important information regarding the size of the structural shift in the data generating process. Thus models that incorporate rapidly this information into future forecasts (such as double differenced devices) are typically the most robust, in the sense that they do not go on to suffer systematic forecast failure after the break.

However, this chapter demonstrates that conventional avenues to robustness are confounded by inaccuracies in measurement. When there are both breaks and measurement errors, the optimal response from a theoretical point of view is to use the information contained in lagged forecast errors, but in a way that gives greater weight to errors that are believed to have been caused by a genuine break. The framework established below shows how the optimal weights depend on the size of any break, relative to the variance of measurement error.

5.1 Forecasting with breaks and measurement error

The forecast-error correction model developed in this chapter builds on an existing literature on intercept correction (IC) models as established error correction devices. Thus this section considers the background of these models, and the extent to which they resolve the problems posed by measurement error and breaks in forecasting.

5.1.1 Forecasting with the conditional expectation

In the situation where only measurement errors affect the model, then the optimal predictor will use all available observations to forecast. Thus for a simple univariate process y_t where:

$$y_t = y^* + \epsilon_t, \quad t = 1, \ldots, T \tag{5.1}$$

with y^* fixed across time, and $\epsilon_t \sim \mathsf{NID}\left(0, \sigma_\epsilon^2\right)$, the forecast of y_t for $t = T+1$ based on the conditional expectation, will be:

$$\mathsf{E}[y_{T+1}|y_1, \ldots, y_T] = \widehat{y}_{T+1|T} = \bar{y} = \frac{1}{T}\sum_{t=1}^{T} y_t$$

such that

$$\begin{aligned}\widehat{\epsilon}_{T+1|T} &= y_{T+1} - \widehat{y}_{T+1|T} \\ &= y^* + \epsilon_{T+1} - y^* - \frac{1}{T}\sum_{t=1}^{T}\epsilon_t\end{aligned}$$

so the forecast error $\widehat{\epsilon}_{T+1|T}$ has expectation $\mathsf{E}[\widehat{\epsilon}_{T+1|T}] = 0$ and mean-squared forecast error (MSFE):

$$\begin{aligned}\mathsf{M}[\widehat{\epsilon}_{T+1|T}] &= \mathsf{E}\left[\left(y_{T+1} - \widehat{y}_{T+1|T}\right)^2\right] \\ &= \sigma_\epsilon^2\left(1 + \frac{1}{T}\right).\end{aligned} \tag{5.2}$$

Since $\widehat{y}_{T+1|T}$ is the conditional expectation, and the data generating process in (5.1) is stationary, it follows from a well-known result (see, e.g. Hamilton (1994, Chapter 4)), that it minimises the mean square forecast error. In-

tuitively this is clear from equation (5.2), since asymptotically, the MSFE converges to σ_ϵ^2, which is the irreducible error generated by the measurement error component ϵ_t. Further, it is also clear that all available observations are used, each with weight $\frac{1}{T}$.

Extending the data generation process to include structural breaks is possible if equation (5.1) becomes:[1]

$$y_t = y^* + \nabla_T \mu + \epsilon_t,$$

where $\nabla_T \mu$ denotes a break of size $\nabla \mu$ that occurs at time T, in similar notation to Chapter 4. In this case, the one-step forecast $\widehat{y}_{T+1|T}$ based on the observations $\{y_t\}_{t=1}^T$ will have an error:

$$\begin{aligned}
\widehat{\epsilon}_{T+1|T} &= y_{T+1} - \widehat{y}_{T+1|T} \\
&= y^* + \nabla_T \mu + \epsilon_{T+1} - \frac{1}{T}\sum_{t=1}^{T} y_t \\
&= \nabla_T \mu - \frac{1}{T}\sum_{t=1}^{T} \epsilon_t
\end{aligned} \tag{5.3}$$

which has expectation

$$\mathsf{E}[\widehat{\epsilon}_{T+1|T}] = \nabla_T \mu,$$

and MSFE

$$\begin{aligned}
\mathsf{M}[\widehat{\epsilon}_{T+1|T}] &= \mathsf{E}\left[(y_{T+1} - \widehat{y}_{T+1|T})^2\right] \\
&= (\nabla_T \mu)^2 + \frac{1}{T}\sigma_\epsilon^2 + \sigma_\epsilon^2 \\
&= (\nabla_T \mu)^2 + \sigma_\epsilon^2 \left(1 + \frac{1}{T}\right)^2.
\end{aligned}$$

As $\nabla_T \mu \neq 0$, $(\nabla_T \mu)^2 > 0$, so the location shift that occurs in y_t and is not accounted for in the forecast $\widehat{y}_{T+1|T}$ induces bias and increases the MSFE.

A more flexible representation discussed by Engle and Hendry (2005)

[1] This notation draws on Engle and Hendry (2005).

allowing for multiple breaks in-sample at $\tau \in T$ yields the model:

$$y_t = y^* + \nabla_\tau \mu + \epsilon_t$$

with

$$\overline{y} = \frac{1}{T}\sum_{t=1}^{T} y_t = \frac{1}{T}\sum_{t=1}^{T} y^* + \frac{1}{T}\sum_{\tau \in T} \nabla_\tau \mu + \frac{1}{T}\sum_{t=1}^{T} \epsilon_t,$$

so that one break at τ^* leads to a forecast error at $T+1$ of

$$\begin{aligned} \widehat{\epsilon}_{T+1|T} &= y^* + \\ & id_\tau \mu + \epsilon_{T+1} - \bar{y} \\ &= \\ id_\tau \mu + \epsilon_{T+1} - \frac{(T-\tau^*)\nabla_{\tau^*} y^*}{T} - \frac{1}{T}\sum_{t=1}^{T} \epsilon_t & \\ &= \epsilon_{T+1} + \frac{\tau^* \nabla_{\tau^*} \mu}{T} - \frac{1}{T}\sum_{t=1}^{T} \epsilon_t, \end{aligned}$$

so that

$$\mathsf{E}[\widehat{\epsilon}_{T+1|T}] = \frac{\tau^* \nabla_{\tau^*} \mu}{T},$$

and

$$\mathsf{M}[\widehat{\epsilon}_{T+1|T}] = \sigma_\epsilon^2 \left(1 + \frac{1}{T}\right) + \left(\frac{\tau^* \nabla_{\tau^*} \mu}{T}\right)^2.$$

The case $\tau^* = T$ is equivalent to a break after forecasting at T, as before, and $\tau^* = 0$ is equivalent to the whole sample covering the break period, with no bias or increase in the MSFE.

When forecasting for period $T+2$ based on $T+1$, the forecast model $\widehat{y}_{T+2|T+1}$ has different properties, depending on whether there is a break or measurement error. Recalling the results above, the forecast error for $T+1$ will be

$$\widehat{\epsilon}_{T+1|T} = y^* - y^* - \frac{1}{T}\sum_{t=1}^{T} \epsilon_t + \epsilon_{T+1}$$

if there is only measurement error, with $\mathsf{E}[\widehat{\epsilon}_{T+1|T}] = 0$. If there is a break

at T, and in the absence of *any* measurement error,

$$\widehat{\epsilon}_{T+1|T} = y^* - y^* + \nabla_T \mu$$

so that $\mathsf{E}[\widehat{\epsilon}_{T+1|T}] = \nabla_T \mu$; allowing for measurement error in addition yields equation (5.3).

If the break still occurs at time T, but now the forecast origin moves one period forward, to $T+1$, then the results are affected by the error ϵ_{T+1}, for both cases of measurement error and breaks. For the former, the forecast error $\widehat{\epsilon}_{T+2|T+1}$ will be:

$$\widehat{\epsilon}_{T+2|T+1} = y^* + \epsilon_{T+2} - y^* - \frac{1}{T+1}\sum_{t=1}^{T+1} \epsilon_t$$

which has expectation $\mathsf{E}[\widehat{\epsilon}_{T+2|T+1}] = 0$ as above; however, if there are breaks alone, in the absence of mismeasurement then

$$\begin{aligned}
\widehat{\epsilon}_{T+2|T+1} &= y^* + \nabla_T \mu - \frac{1}{T+1}\sum_{t=1}^{T+1} y_t \\
&= y^* + \nabla_T \mu - \frac{1}{T+1}\left((T+1)y^* + \nabla_T \mu\right) \\
&= \nabla_T \mu \left(1 - \frac{1}{T+1}\right) \\
&= \nabla_T \mu \left(\frac{T}{T+1}\right),
\end{aligned}$$

with

$$\mathsf{E}[\widehat{\epsilon}_{T+2|T+1}] = \nabla_T \mu \left(\frac{T}{T+1}\right)$$

and

$$\mathsf{M}[\widehat{\epsilon}_{T+2|T+1}] = \left(\nabla_T \mu \left(\frac{T}{T+1}\right)\right)^2.$$

Thus the bias in the forecast when forecasting from $T+1$ is smaller than at T, when the break is pre-forecasting; this follows from the fact that at $T+1$, the information set $\mathcal{I}_{T+1} = (y_1, \ldots, y_{T+1})$ includes some post-break

information, which is used in generating the forecast for $T+2$.[2]

Looking at the problem from an alternative perspective, when there are *no* breaks but there is some measurement error, then the fact that 'contaminated' (i.e. post-break) observations are included in the information set can lead to problems. For example, if only measurement error affects observations at $T+1$ and $T+2$ then:

$$\widehat{y}_{T+2|T+1} = \frac{1}{T+1}\sum_{t=1}^{T+1} y_t = y^* + \frac{\epsilon_{T+1}}{T+1}$$

so that

$$\widehat{\epsilon}_{T+2|T+1} = y^* + \epsilon_{T+2} - y^* - \frac{\epsilon_{T+1}}{T+1}$$

with

$$\mathsf{E}[\widehat{\epsilon}_{T+2|T+1}] = 0$$

and

$$\begin{aligned}\mathsf{M}[\widehat{\epsilon}_{T+2|T+1}] &= \sigma_\epsilon^2\left(1 - \frac{1}{(T+1)^2}\right) \\ &= \sigma_\epsilon^2\left(\frac{T(T+2)}{(T+1)^2}\right).\end{aligned}$$

Clearly for large T the MSFE will be relatively small, for a given σ_ϵ^2, but a significant problem emerges when there may be both breaks and measurement error. In this case, the choice of whether or not to incorporate the observation at $T+1$ when forecasting one step ahead to $T+2$ can have a large effect on forecast performance, and should in some way be dictated by the relative likelihood of one effect over the other. This issue lies at the heart of the remainder of this chapter.

[2]For the general one-step forecast for $T+k+1$ at $T+k$, for any $k>0$, the error is:

$$\widehat{\epsilon}_{T+k+1|T+k} = \nabla_T \mu\left(\frac{T}{T+k}\right),$$

which tends to 0 as $k \to \infty$.

5.1.2 Robust forecasting models

When there is a structural break, the previous section suggested that the most valuable observations are those in the post-break sample, since pre-break data will be uninformative. This suggests that one obvious candidate for a 'robust' model – namely one which does not suffer systematic forecast failure when there is a shift – denoted $\widetilde{y}_{T+1|T}$, should use *only* the latest observation:

$$\widetilde{y}_{T+1|T} = y_T.$$

In the absence of measurement error, this delivers the best (minimum MSFE) forecast, since although for a break at T, the forecast for $T+1$ given T will be biased:

$$\begin{aligned}\widetilde{\epsilon}_{T+1|T} &= y_{T+1} - \widetilde{y}_{T+1|T} \\ &= y^* + \nabla_T \mu - y^* \\ &= \nabla_T \mu\end{aligned}$$

with MSFE

$$\mathsf{M}[\widetilde{\epsilon}_{T+1|T}] = (\nabla_T \mu)^2,$$

one period *after* this:

$$\widetilde{\epsilon}_{T+2|T+1} = y_{T+2} - \widetilde{y}_{T+2|T+1} = 0$$

and $\mathsf{M}[\widetilde{\epsilon}_{T+2|T+1}] = 0$. The success of this model derives from the fact that y_T *is* the data generating process for y up to the point T, and so any differences between y_{T+1} and y_T will reflect changes between these two points alone; as such, no estimation is necessary. As Clements and Hendry observe in the context of using similar robust models:

> '... without the forecaster knowing the causal variables or the structure of the economy, or whether there have been any structural breaks or shifts, and without any estimation needed ... [this] reflects all the effects in the DGP.'
>
> Clements and Hendry (2005, p.739)

However, the choice of a robust model is not free of penalties: if there is measurement error, then:

$$\widetilde{\epsilon}_{t+1|t} = y^* + \epsilon_{t+1} - y^* - \epsilon_t$$

so

$$\mathsf{M}[\widetilde{\epsilon}_{t+1|t}] = 2\sigma_\epsilon^2.$$

Compared to the (break-free) MSFE in the conditional expectation forecast model of $\sigma_\epsilon^2(1+T^{-1})$ this will be larger for $T > 1$, and in the case of a break at T:[3]

$$\begin{aligned}\mathsf{M}_\mathsf{A} &= \mathsf{M}[\widehat{\epsilon}_{T+2|T+1}] = \sigma_\epsilon^2\left(1+\frac{1}{T+1}\right) + \left(\nabla_T\mu\frac{T}{T+1}\right)^2 \\ \mathsf{M}_\mathsf{B} &= \mathsf{M}[\widetilde{\epsilon}_{T+2|T+1}] = 2\sigma_\epsilon^2\end{aligned}$$

and $\mathsf{M}_\mathsf{B} < \mathsf{M}_\mathsf{A}$ if:

$$\begin{aligned}\sigma_\epsilon^2\left(1+\frac{1}{T+1}\right) + \left(\nabla_T\mu\frac{T}{T+1}\right)^2 &> 2\sigma_\epsilon^2 \\ (\nabla_T\mu)^2 &> \left(\frac{T+1}{T}\right)\sigma_\epsilon^2\end{aligned}$$

As T becomes large, this amounts to the (absolute) size of the break being larger than one standard deviation of the measurement error ϵ_t; clearly in the absence of breaks, this condition is never satisfied.

5.1.3 Forecasting in dynamic systems

The analysis thus far has considered static models in which y_t does not depend on past values of y. Extending this to consider dynamic models introduces a new dimension to the problem, and sets the foundation for forecast error correction models. Consider the process:

$$y_t^* = \phi y_{t-1}^* + \eta_t,$$

[3] Denoting the MSFE of the conditional-expectation model by M_A and that of the IC model as M_B.

where y_t^* is the 'true' value of y_t, $|\phi| < 1$ and $\eta_t \sim \mathsf{NID}\left(0, \sigma_\eta^2\right)$. If y_t^* is only observed with measurement error (so that the truth is a latent variable), such that the observation

$$y_t = y_t^* + \epsilon_t$$

where $\epsilon_t \sim \mathsf{NID}\left(0, \sigma_\epsilon^2\right)$ is independent of η_t, then substitution into the preceding equation yields

$$(y_t - \epsilon_t) = \phi(y_{t-1} - \epsilon_{t-1}) + \eta_t$$

which, letting $v_t = \epsilon_t - \phi\epsilon_{t-1} + \eta_t$, can be written as:

$$y_t = \phi y_{t-1} + v_t.$$

Although $\mathsf{E}[v_t] = 0$, $\mathsf{V}[v_t] = \sigma_\eta^2 + (1+\phi^2)\sigma_\epsilon^2$, so measurement error inflates the variance due to both the contemporaneous and lagged observations of y; further, it induces a negative moving average structure in the error term, such that:

$$\mathsf{Cor}[v_t, v_{t-j}] = \frac{-\phi\sigma_\epsilon^2}{\sigma_\eta^2 + (1+\phi^2)\sigma_\epsilon^2} \quad j = 1,$$

though higher-order correlations (i.e. $j > 1$) are zero. This in turn renders the estimate $\widehat{\phi}$, if measurement error affects the sample period, both biased and inconsistent.

Assuming that only the post-estimation period is affected by measurement errors and breaks,[4] a one-step forecast based on the conditional expectation at time T with known ϕ will be:

$$\widehat{y}_{T+1|T} = \phi y_T, \tag{5.4}$$

so:

$$\begin{aligned}\widehat{\eta}_{T+1|T} &= y_{T+1}^* - \phi y_T \\ &= \eta_{T+1} + \epsilon_{T+1}\end{aligned}$$

where $\mathsf{E}[\widehat{\eta}_{T+1|T}] = 0$, and $\mathsf{M}[\widehat{\eta}_{T+1|T}] = \sigma_\eta^2 + \sigma_\epsilon^2$. One step beyond this,

[4]So for $t \leq T$, $y_t = y_t^*$ and $y_t = y_t^* + \epsilon_t$ thereafter.

still maintaining the assumption of no breaks, and the forecast error now includes measurement error carried over from the previous period, $T+1$:

$$\begin{aligned}\widehat{\eta}_{T+2|T+1} &= y^*_{T+2} - \phi y_{T+1} \\ &= \eta_{T+2} + \epsilon_{T+2} - \phi\epsilon_{T+1}\end{aligned}$$

which has MSFE

$$\mathsf{M}[\widehat{\eta}_{T+2|T+1}] = \sigma^2_\eta + (1+\phi^2)\sigma^2_\epsilon = \mathsf{V}[v_t].$$

Thus in a dynamic system, the conditional expectation forecast behaves in a similar way to its static counterpart: for $\sigma^2_\epsilon = 0$, $\mathsf{E}[\widehat{\eta}_{T+1|T}] = 0$ and $\mathsf{M}[\widehat{\eta}_{T+1|T}] = \sigma^2_\eta$, which is the irreducible error from the data generating process.

As might be expected, breaks can cause problems; for a shift in the equilibrium mean at T from zero to $\nabla_T\mu$, the DGP at $T+1$ will be:

$$y_{T+1} = \phi y_T + \nabla_T\mu + \eta_{T+1} + \epsilon_{T+1}.$$

Since the forecast model ignores the break, it is given by:

$$\widehat{y}_{T+1|T} = \phi y_T$$

which has a corresponding error:

$$\widehat{\eta}_{T+1|T} = \nabla_T\mu + \eta_{T+1} + \epsilon_{T+1},$$

which has expectation

$$\mathsf{E}[\widehat{\eta}_{T+1|T}] = \nabla_T\mu,$$

and MSFE

$$\mathsf{M}[\widehat{\eta}_{T+1|T}] = (\nabla_T\mu)^2 + \sigma^2_\eta + \sigma^2_\epsilon.$$

Further, assuming that parameter estimation is unaffected by measurement error, so that related uncertainty can be ignored, the forecast error for $T+$

$h+1$ at $T+h$ will also be:

$$\widehat{\eta}_{T+h+1|T+h} = \nabla_T \mu + \eta_{T+h+1} + \epsilon_{T+h+1} - \phi\epsilon_{T+h},$$

such that

$$\mathsf{E}[\widehat{\eta}_{T+h+1|T+h}] = \nabla_T \mu, \tag{5.5}$$

with a similar MSFE as before, this time inflated by the $\phi\epsilon_{T+h}$ term. Thus the effect on the forecast model, of missing a break in the DGP is systematic bias, as several previous studies have established (see, for example, Hendry (2006)).

An interesting comparison to make in the dynamic DGP case is that of the forecast errors of a one-step model for $T+2|T+1$, and a two-step model $T+2|T$. Considering the latter, note that:

$$\widehat{y}_{T+2|T} = \phi\widehat{y}_{T+1|T} = \phi^2 y_T.$$

However, since:

$$y_{T+2} = \phi^2 y_T + (1+\phi)\nabla_T \mu + \sum_{i=0}^{1} \phi^i \eta_{T+2-i} + \sum_{i=0}^{1} \phi^i (\epsilon_{T+2-i} - \phi\epsilon_{T+1-i})$$

it follows that:

$$\begin{aligned}\widehat{\eta}_{T+2|T} &= y_{T+2} - \widehat{y}_{T+2|T} \\ &= (1+\phi)\nabla_T \mu + \sum_{i=0}^{1} \phi^i \eta_{T+2-i} + \sum_{i=0}^{1} \phi^i (\epsilon_{T+2-i} - \phi\epsilon_{T+1-i})\end{aligned}$$

which has expectation

$$\mathsf{E}[\widehat{\eta}_{T+2|T}] = (1+\phi)\nabla_T \mu,$$

which is absolutely greater than (5.5) if $\nabla_T \mu \neq 0$. This is consistent with intuition: if the break in the mean happens after *both* the one-step and two-step forecasts are produced, then both will be affected, the latter more severely on account of its recursive nature. If, on the other hand, either measurement error or breaks affect the DGP *in between* the one- and two-

step forecasts being produced, then the effect is ambiguous: if there is a mean shift, the one-step forecast will dominate, but a large measurement outlier at $T+1$ will act like a break in the one-step forecast, but be ignored in the two-step case.

Lying at the heart of this result is the fact that observing *one* forecast error is insufficient when making a judgment on whether forecast failure is due to breaks or mis-measurement. Thus an appropriate direction for forecast models to pursue should be one using the past few forecast errors in some kind of weighted combination, and so the next section considers a possible example of this.

5.2 Forecast-error correction models

A simple example of a forecast-error correction mechanism is the intercept correction (IC) model studied in detail by, e.g. Clements and Hendry (1998). The central principle is that ICs should set a forecast model 'back on track' if there is a location shift in the mean, and one possible way to achieve this is through the direct insertion of a lagged forecast error. In a measurement error-free context, an IC can be demonstrated with a straightforward dynamic model:

$$y_t = \nabla_T \mu + \phi y_{t-1} + \eta_t,$$

where ϕ lies within the unit circle, η_t behaves as before, and $\nabla_T \mu$ denotes a change in the equilibrium mean of y_t, from zero up to and including T to $\nabla_T \mu$ afterward. By expressing this shift in mean as:

$$\nabla_T \mu = \nabla_T \mu,$$

an equivalent interpretation of the DGP is:

$$y_t = \phi y_{t-1} + \nabla_T \mu + \eta_t$$

where the intuition behind the process being in equilibrium-correction form is clearer. Since the $\nabla_T \mu$ component only starts to affect y_t from $T+1$ onwards, the first non-trivial application of an IC is for the immediate

aftermath of the break. As established in the previous section, at $T+1$ the error arising from a conditional expectation forecast, $\widehat{\eta}_{T+1|T}$ will be:

$$\begin{aligned}\widehat{\eta}_{T+1|T} &= y_{T+1} - \widehat{y}_{T+1|T} \\ &= \nabla_T \mu + \eta_{T+1}\end{aligned}$$

if parameter estimation uncertainty is ignored. In order to set future forecasts back on track (since, as discussed above, the break will cause systematic bias if left uncorrected), a possible route is to insert $\widehat{\eta}_{T+1|T}$ directly into the forecast for $T+2|T$:

$$\begin{aligned}\widetilde{y}_{T+2|T+1} &= \phi y_{T+1} + \widehat{\eta}_{T+1|T} \\ &= \phi y_{T+1} + \nabla_T \mu + \eta_{T+1}.\end{aligned}$$

This leads to a forecast error of:

$$\widetilde{\eta}_{T+2|T+1} = \eta_{T+2} - \eta_{T+1}$$

which has zero expectation, although its MSFE $\mathsf{M}[\widetilde{\eta}_{T+2|T+1}] = 2\sigma_\eta^2$ is twice that of the standard conditional expectation forecast, which is the irreducible DGP innovation error.

Allowing for measurement error yields a similar result, although the MSFE is inflated, since:

$$\widehat{\eta}_{T+1|T} = \nabla_T \mu + \eta_{T+1} + \epsilon_{T+1} - \phi\epsilon_T \tag{5.6}$$

so

$$\widetilde{y}_{T+2|T+1} = \phi y_{T+1} + \widehat{\eta}_{T+1|T},$$

with a resulting error of

$$\widetilde{\eta}_{T+2|T+1} = \Delta\eta_{T+2} + \Delta\epsilon_{T+2} - \phi\Delta\epsilon_{T+1},$$

which has zero expectation but MSFE

$$\mathsf{M}[\widetilde{\eta}_{T+2|T+1}] = 2\sigma_\eta^2 + \sigma_\epsilon^2(2 + \phi(1+2\phi)). \tag{5.7}$$

From this equation, it is easy to see why the presence of measurement error could undermine efforts to use ICs as a route to robust forecasts: compared to the variance of a correction-free forecast from the dynamic model examined in the previous section, the MSFE of $\widetilde{\eta}_{T+2|T+1}$ not only suffers from inflated innovation errors ($2\sigma_\eta^2$), but also a significantly amplified measurement error component. From the perspective of correcting for a location shift, though, the success of $\widetilde{y}_{T+2|T+1}$ is evident, through the fact that no $\nabla_T \mu$ term appears in the forecast error; this highlights the trade-off that represents the crux of this chapter.

However, the problem does not seem insurmountable; in the same way that a ship navigating at night via the signals from lighthouses might be able to establish exactly its position when observing *two* separate beams, rather than one (largely uninformative) point, a logical way to distinguish between stochastic measurement error (which might affect one period, but not be carried over) and a structural break (which would *persist*) is to use further past forecast errors. An obvious candidate model with two lags, denoted IC(2), is given by:

$$\widetilde{y}_{t+j+1|t+j} = \phi y_{t+j} + \omega_1 \widehat{\eta}_{t+j|t+j-1} + \omega_2 \widehat{\eta}_{t+j-1|t+j-2}.$$

For now, it is assumed that the weights ω_i on the lagged errors are equal, so that $\omega_1 = \omega_2$, and $\sum_{i=1}^{2} \omega_i = 1$, so that in the IC(2) case, $\omega_i = \frac{1}{2}$; relaxing this assumption is the subject of the following section.

The lagged forecast error, is:

$$\widehat{\eta}_{t+i+1|t+i} = y_{t+i+1} - \widehat{y}_{t+i+1|t+i}$$

where $\widehat{y}_{t+i+1|t+i}$ is the conventional conditional expectation forecast. For an IC(2), for a mean shift at $T = t + j - 2$, the one-step forecast for $T + 3$ will be:

$$\widetilde{y}_{T+3|T+2} = \phi y_{T+2} + \frac{1}{2}\widehat{\eta}_{T+2|T+1} + \frac{1}{2}\widehat{\eta}_{T+1|T}. \tag{5.8}$$

This delivers a forecast error $\widetilde{\eta}_{T+3|T+2}$ of:

$$\begin{aligned}\widetilde{\eta}_{T+3|T+2} =& \phi y_{T+2} + \nabla_T \mu + \eta_{T+3} + \epsilon_{T+3} - \phi\epsilon_{T+2} \\ &- \phi y_{T+2} - \frac{1}{2}\left(\eta_{T+2} + \epsilon_{T+2} - \phi\epsilon_{T+1}\right) - \frac{1}{2}\nabla_T \mu \\ &- \frac{1}{2}\left(\eta_{T+1} + \epsilon_{T+1} - \phi\epsilon_T\right) - \frac{1}{2}\nabla_T \mu\end{aligned}$$

which has expectation $\mathsf{E}[\widetilde{\eta}_{T+3|T+2}] = 0$, since the $\nabla_T \mu$ term that enters through the y_{T+3} component is exactly offset by the two $\frac{1}{2}\nabla_T \mu$ terms that enter through the lagged forecast errors. Further,

$$\mathsf{M}[\widetilde{\eta}_{T+3|T+2}] = \frac{3\sigma_\eta^2}{2} + \sigma_\epsilon^2\left(\frac{3}{2} + \frac{\phi}{2} + \frac{3\phi^2}{2}\right),$$

which is always lower than the IC(1) MSFE in (5.7). The intuition behind this result seems to lie with the fact that the individual weight on each lagged forecast error is strictly less than one, and consequently their joint contribution to the MSFE is weighted *down*; thus the vulnerability of the forecast model to a *particular measurement error* is reduced.

As a direct consequence of insulation from measurement errors, though, robustness to breaks is affected: for example, for a break occurring at $T+1$, rather than T, the forecast error above has expectation:

$$\mathsf{E}[\widetilde{\eta}_{T+3|T+2}] = \frac{1}{2}\nabla_{T+1}\mu \neq 0,$$

and the MSFE increases. Thus the trade-off between robustness to breaks and measurement error is still present. However, expressing the problems in terms of a choice of weights, ω_i, frames the problem in a more tractable way.

5.3 Optimal Weights for an IC(2) model

Allowing the weights $\boldsymbol{\omega}$ to vary with the parameters in the data generating process yields the theoretical 'optimal' weights, which minimise the MSFE $\mathsf{M}[\widetilde{\eta}_{T+3|T+2}]$. Denoting ω_1 and ω_2 as the weights on $\widehat{\eta}_{T+2|T+1}$ and $\widehat{\eta}_{T+1|T}$

respectively, then (5.8) becomes:

$$\widetilde{y}_{T+3|T+2} = \phi y_{T+2} + \omega_1 \widehat{\eta}_{T+2|T+1} + \omega_2 \widehat{\eta}_{T+1|T}.$$

Thus if a structural break occurs at point T, then the forecast error $\widetilde{\eta}_{T+3|T+2}$ will be given by:

$$\begin{aligned}\widetilde{\eta}_{T+3|T+2} =& \nabla_T \mu(1-\omega_1-\omega_2) + \eta_{T+3} + \epsilon_{T+3} - \phi\epsilon_{T+2} \\ &- \omega_1(\eta_{T+2} + \epsilon_{T+2} - \phi\epsilon_{T+1}) \\ &- \omega_2(\eta_{T+1} + \epsilon_{T+1} - \phi\epsilon_T).\end{aligned}$$

This delivers a forecast bias of:

$$\mathsf{E}[\widetilde{\eta}_{T+3|T+2}] = \nabla_T \mu(1-\omega_1-\omega_2)$$

and a MSFE of:

$$\begin{aligned}\mathsf{M}[\widetilde{\eta}_{T+3|T+2}] =& (\nabla_T \mu)^2(1-\omega_1-\omega_2)^2 + \sigma_\eta^2(1+\omega_1^2+\omega_2^2) \\ &+ \sigma_\epsilon^2(1+\phi^2)(1+\omega_1^2+\omega_2^2) - 2\sigma_\epsilon^2\omega_1\omega_2\phi \\ &+ 2\sigma_\epsilon^2\phi\omega_1. \end{aligned} \tag{5.9}$$

In order to find the optimal weights ω_1^* and ω_2^*, the first step is to differentiate (5.9) with respect to ω_1 and ω_2. This yields a set of simultaneous equations:

$$\begin{pmatrix} \Omega & -\Phi \\ -\Phi & \Omega \end{pmatrix} \begin{pmatrix} \omega_1^* \\ \omega_2^* \end{pmatrix} = \begin{pmatrix} -\Phi \\ \left(\frac{\nabla_T \mu}{\sigma_\eta}\right)^2 \end{pmatrix}, \tag{5.10}$$

where:

$$\Omega = \left(\frac{\nabla_T \mu}{\sigma_\eta}\right)^2 + 1 + \frac{\sigma_\epsilon^2}{\sigma_\eta^2}(1+\phi^2),$$

and

$$\Phi = \frac{\sigma_\epsilon^2}{\sigma_\eta^2}\phi - \left(\frac{\nabla_T \mu}{\sigma_\eta}\right)^2,$$

where $\frac{\sigma_\epsilon^2}{\sigma_\eta^2} = q$ is the noise-to-signal ratio.

Since the Hessian of second and cross- partial derivatives has determinant

$$\begin{vmatrix} \Omega & -\Phi \\ -\Phi & \Omega \end{vmatrix} = \Omega^2 - \Phi^2 > 0$$

it follows that the optimal weights $\boldsymbol{\omega}^*$ minimise the MSFE. By Cramer's rule, these are given by:

$$\omega_1^* = \frac{\left(\frac{\nabla_T \mu}{\sigma_\eta}\right)^2 (\Omega + \Phi) - q\phi\Omega}{\Omega^2 - \Phi^2} \tag{5.11a}$$

and

$$\omega_2^* = \frac{\left(\frac{\nabla_T \mu}{\sigma_\eta}\right)^2 (\Omega + \Phi) - q\phi\Phi}{\Omega^2 - \Phi^2}. \tag{5.11b}$$

Thus the optimal weight depends on the size of any mean shift, and in this case it is assumed that such a shift takes place in T and lasts for at least three periods. In this sense, the intuition behind each weight is clear: when there is no mean shift at all, and no error in measurement, then the optimal weights $\boldsymbol{\omega}^* = \mathbf{0}$, since the lagged forecast errors contain *no* relevant information about any structural shifts that could have taken place. When there are breaks alone, then this is not the case, since the lagged errors yield an indication of its size, although the inclusion of lagged DGP innovation errors leads to a concomitant increase in forecast uncertainty. Thus the optimal weights $\boldsymbol{\omega}^*$ in this case depend on the size of the break alone:

$$\omega_1^* = \omega_2^* = \frac{\left(\frac{\nabla \mu}{\sigma_\eta}\right)^2}{2\left(\frac{\nabla \mu}{\sigma_\eta}\right)^2 + 1}. \quad \sigma_\epsilon^2 < \infty \tag{5.12}$$

5.3.1 Optimal Weights: limiting cases

Although a *particular* weight will depend on the specific parameters ϕ, σ_ϵ^2, σ_η^2 and $\nabla_T \mu$, it is worth considering how $\boldsymbol{\omega}^*$ behave as different components of Ω, Φ and $(\nabla_T \mu / \sigma_\eta^2)$ become very small or very large. As the standardised (squared) size of the mean shift becomes large, so $\left(\frac{\nabla_T \mu}{\sigma_\eta}\right)^2 \to \infty$, then $\omega_1^* = \omega_2^* \to \frac{1}{2}$, since the 'value' of correcting for the mean shift outweighs the cost of additional DGP noise, and straightforward calculations yield the

condition that $\omega_i^* \in \left[0, \frac{1}{2}\right)$ for $i = 1, 2$.[5]

As might be expected, letting the measurement error variance σ_ϵ^2 become large leads to an opposite conclusion, since as $\sigma_\epsilon^2 \to \infty$:

$$\omega_1^* \to -\frac{\phi(1+\phi^2)}{1+\phi^2+\phi^4} \tag{5.13a}$$

$$\omega_2^* \to -\frac{\phi^2}{1+\phi^2+\phi^4}. \tag{5.13b}$$

The dependence on ϕ is such that for $\phi = 0$, $\omega_1^* = \omega_2^* \to 0$, but as $\phi \to 1$, $\omega_1^* \to -\frac{2}{3}$ and $\omega_2^* \to -\frac{1}{3}$. Although it might seem counter-intuitive to place *any* weight on lagged forecast errors even if there has been no mean shift, the explanation lies in the fact that the presence of measurement error induces a negative moving average in the error, as discussed in Section 5.1.3. The MA structure yields a non-zero covariance between the two lagged forecast errors, under the conditions that $\phi \neq 0$ and $\sigma_\epsilon^2 > 0$, and in the case where $\phi > 0$, negative weights on the errors can *reduce* the MSFE relative to the baseline case without any intercept correction.

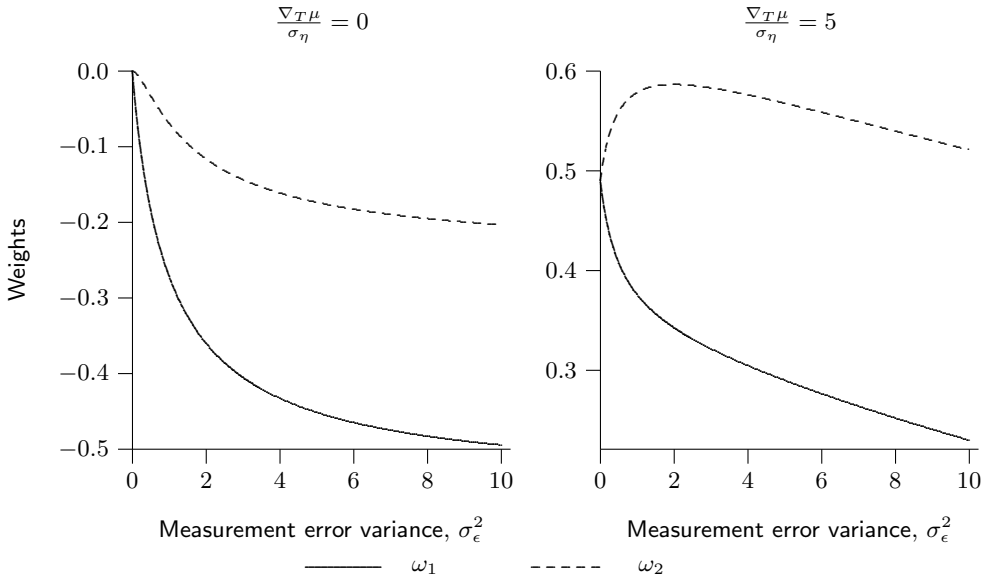
By allowing for different degrees of measurement error noise, and different break sizes, Figure 5.1 shows the optimal weights across a range of the parameter values. In the limit, as $\left(\frac{\nabla_T \mu}{\sigma_\eta}\right)^2 \to \infty$, for $\sigma_\epsilon^2 = 0$, the weights for both ω_1^* and ω_2^* tend to $\frac{1}{2}$, as noted above; as σ_ϵ^2 increases, though, both tend to the limiting case described by (5.13a) and (5.13b), albeit from a higher starting point for larger mean shifts. Of particular interest in the right-hand panel of Figure 5.1 is the behaviour of ω_2^*, where the weight is *increasing* in measurement error noise for small values of σ_ϵ^2, before dropping as the variance continues to increase. The rationale for this lies in the fact that the weights in the system (5.10) are 'asymmetric' since the first row of the system includes the term $\phi\frac{\sigma_\epsilon^2}{\sigma_\eta^2}$, which is the autocovariance of the error term

[5] Let

$$A\left(\frac{\nabla_T \mu}{\sigma_\eta}\right)^2 = 2\left(\frac{\nabla_T \mu}{\sigma_\eta}\right)^2 + 1.$$

Then, since $\left(\frac{\nabla_T \mu}{\sigma_\eta}\right)^2 \geq 0$ and $\sigma_\eta^2 < \infty$, it follows that $A \in (2, \infty)$, so $\omega_i^* < \frac{1}{2}$, $i = 1, 2$, $\forall \left(\frac{\nabla_T \mu}{\sigma_\eta}\right)^2$.

Figure 5.1: OPTIMAL IC(2) WEIGHTS WITH INCREASING MEASUREMENT UNCERTAINTY



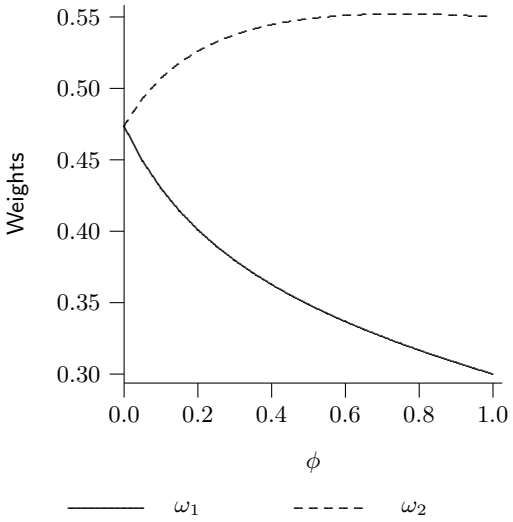
NOTES: In each case, the optimal weights ω_1^* and ω_2^* are calculated for $\phi = 0.6$, $\sigma_\eta^2 = 1$. On the left, there is no break in the mean of y_t, and on the right, there is a five standard deviation shift in the mean.

v_t, scaled by the DGP innovation variance.[6] As a result, an increase in σ_ϵ^2 from zero introduces this covariance term in the optimal weight calculations, and since the term enters via the inclusion of the *most recent* lagged forecast error, the optimal weight ω_1^* falls; in contrast, the effect on ω_2^* is positive for small values of σ_ϵ^2 before the signal of the mean shift is lost in the noise of the composite error term v_t.

Since the covariance between v_{t+k} and v_{t+k-1} (and therefore $\widehat{\eta}_{t+k}$ and $\widehat{\eta}_{t+k-1}$) appears to be important in explaining several underlying results, it is instructive to conclude the examination of limiting cases by changing the size of the autoregressive parameter, ϕ, in the DGP. Figure 5.2 provides a graphical demonstration for given values of the relevant parameters, but it is helpful to isolate the effect of specific elements by looking at the algebraic expressions when $\phi = 0$ and $\phi = 1$.

[6]The source of this asymmetry can be found in the third row of equation (5.9), which comprises a term in ω_1 alone.

Figure 5.2: OPTIMAL IC(2) WEIGHTS ACROSS ϕ



NOTES: As in Figure 5.1, optimal weights ω_1 and ω_2 are calculated, with $\frac{\nabla_T \mu}{\sigma_\eta} = 5$, $\sigma_\eta^2 = 1$ and $\sigma_\epsilon^2 = 1$.

Taking the former case, the optimal weights are equal, and given by:

$$\omega_1^*\Big|_{\phi=0} = w_2^*\Big|_{\phi=0} = \frac{\left(\frac{\nabla_T \mu}{\sigma_\eta}\right)^2}{2\left(\frac{\nabla_T \mu}{\sigma_\eta}\right)^2 + (1+q)},$$

where the similarity to the case of no measurement error in equation (5.12) is obvious, and explained by the fact that, for $\phi = 0$, the problem collapses to a 'conventional' one of trading off increased mean squared forecast error due to bias, against greater uncertainty.

When $\phi = 1$, the optimal weights are a little more involved:

$$\begin{aligned}\omega_1^*\Big|_{\phi=1} &= \frac{\left(\frac{\nabla_T \mu}{\sigma_\eta}\right)^2 (1+3q) - q\left(1+2q+\left(\frac{\nabla_T \mu}{\sigma_\eta}\right)\right)}{2\left(\frac{\nabla_T \mu}{\sigma_\eta}\right)^2 (1+3q) + (1+3q)(1+q)} \\ &= \omega_1^*\Big|_{\phi=0} - \frac{q\left(1+2q+\left(\frac{\nabla_T \mu}{\sigma_\eta}\right)^2\right)}{2\left(\frac{\nabla_T \mu}{\sigma_\eta}\right)^2 (1+3q) + (1+3q)(1+q)} \qquad (5.14a)\end{aligned}$$

and similarly:

$$\omega_2^*\Big|_{\phi=1} = \omega_2^*\Big|_{\phi=0} - \frac{q^2 - q\left(\frac{\nabla_T \mu}{\sigma_\eta}\right)^2}{2\left(\frac{\nabla_T \mu}{\sigma_\eta}\right)^2 (1+3q) + (1+3q)(1+q)} \tag{5.14b}$$

Thus whilst the second term in equation (5.14a) is never positive, so

$$\omega_1^*\big|_{\phi=1} \leq \omega_1^*\big|_{\phi=0},$$

the same is not true for w_2^*: for a sufficiently large break,

$$\omega_2^*\big|_{\phi=1} > \omega_2^*\big|_{\phi=0},$$

which once again stems from the fact that the optimal weights depend on the covariance between the two lagged forecast errors.

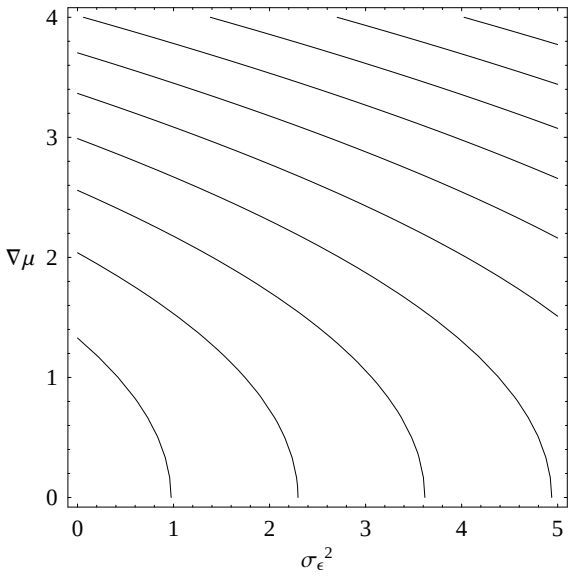
5.3.2 Strategies for robust forecasts

It is possible to capture the MSFE trade-off between the size of measurement error noise and the magnitude of a mean shift, assuming that for any particular combination of the two, the optimal weights are selected according to (5.11a) and (5.11b). The motivation for examining this lies in the fact that it provides an indication of the optimal approach to building 'robust' forecast models, where the objective of such a model is to deliver a small MSFE following a structural shift in the DGP. Specifically, the nature of the trade-off can reveal whether the best response to a combination of measurement error and breaks is to seek to 'insure' a model *completely* against one over the other.

Thus Figures 5.3 and 5.4 plot the MSFE 'contours' for different combinations of mean shift $\left(\frac{\nabla_T \mu}{\sigma_\eta}\right)$ and measurement error variance σ_ϵ^2 in two cases: first, when the forecast model is simply the conditional expectation (*viz.* equation (5.4)), and secondly, when an IC(2) with optimal weights is used. Each solid contour line relates to a particular MSFE, which is increasing as lines move away from the origin at (0,0).

One of the most important observations that can be made with regard to

Figure 5.3: VARIANCE-MEAN SHIFT TRADE-OFF: CONDITIONAL EXPECTATION FORECAST



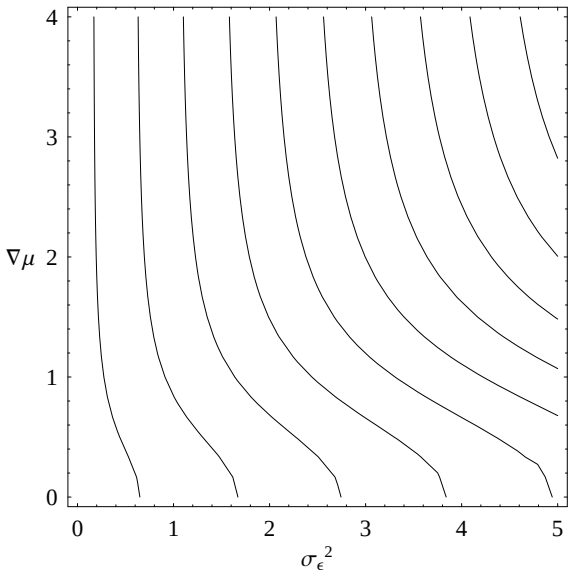
NOTES: Contour lines represent the MSFE from a conditional expectation forecast model of the kind in equation (5.4), for different sizes of structural break ($\nabla\mu$), or measurement error (σ_ϵ^2) in the data generating process. Each line is the locus of points for which the MSFE is the same, with the forecast error increasing as the lines move further away from the origin at $0, 0$.

Figure 5.3 is that the contour lines are strictly concave for all combinations of mean shift and error variance. The intuition behind this is relatively clear, since both elements induce forecast failure, either through bias in the case of the former, or increased variance in the case of the latter, and these elements operate independently of one another. Following from this, a key implication regarding forecasting strategies is that a robust forecast model should incorporate an element of insurance against *both* sources of error.

In this spirit, the contours of the IC(2) function, which is designed to provide exactly this kind of robustness, behave in a radically different way, as Figure 5.4: for sufficiently large mean shifts,[7] the contours are convex,

[7]Greater than approximately half of one standard deviation of the DGP innovation error.

Figure 5.4: VARIANCE-MEAN SHIFT TRADE-OFF: IC(2) FORECAST



NOTES: Contour lines represent the MSFE from a varying-weight IC(2) model, as explained in the text. For different sizes of structural break ($\nabla\mu$), or measurement error (σ_ϵ^2) in the data generating process, the IC(2) model calculates the optimal weight to attach to lagged forecast errors. Each line is the locus of points for which the MSFE is the same, assuming that for each point the optimal weights are being used, with the forecast error increasing as the lines move further away from the origin at $0, 0$.

but whilst the MSFE is strictly increasing in the error variance, it is simply non-decreasing in the size of the mean shift. Thus asymptotically as the size of a break becomes very large, the marginal effect on the MSFE tends to zero. As a mechanism of achieving a robust forecast, the IC(2) function clearly succeeds in this particular application; further, the contour approach to analysing the effect of robust devices offers a useful platform within which to compare differing approaches.

5.4 Concluding comments

The foregoing comparison of optimal weights in different circumstances is of use in understanding some of the principles that underlie a robust forecast-

ing strategy. Since the role of the covariance between forecast errors $\widehat{\eta}_{t+k}$ and $\widehat{\eta}_{t+k-1}$, which arises due to measurement error, is significant in explaining the behaviour of the weights in several cases, this would suggest that there are fundamental differences between the optimal approach to the DGP studied here, and the literature on models with only breaks and white noise error terms. Owing to the richer covariance structure in the former, optimal forecast weights can be different, perhaps explaining why 'conventional' robust devices can encounter problems in empirical forecasting exercises, when measurement error is likely to be present.

Therefore, even though the varying-weight model discussed here is not effective in practice, as it requires knowledge of the variance of measurement error, the DGP innovation variance and the magnitude of any break that occurs, it nonetheless provides a useful insight into the behaviour of existing models, and an avenue for future research into deriving consistent estimates of the weights.

Considering the wider question of how models should respond to breaks, the message from this chapter is that the information that is generated when a break occurs can be valuable. Thus chapter 6 considers how different forecast models might be used in a real-world forecasting problem.

Chapter 6

Learning and Forecasting: UKM1 Revisited

> *Mr James Callaghan, Leader of the Opposition* – . . . will she tell us clearly whether increases in wages are a cause of inflation or not?
>
> *Mrs Thatcher* – Over a period the cause of increased inflation is increases in the money supply. Within money supply, there will be a different distribution both between the public sector and the private sector and within those sectors there will be increases in pay within the general money supply well beyond what are warranted, and they may come through in increases in particular products which will not necessarily affect the general price level.
>
> *Mr James Callaghan* – May I thank her for that reply and say that I did not understand a word of it.
>
> HANSARD, 3 JULY 1980
> *Quoted in Kitson and Michie (2000)*

Structural breaks seem to occur frequently in economic time series, and as earlier chapters demonstrated, they present a serious threat to the accuracy of a forecast model. An important step towards developing a forecast strategy robust to breaks is to understand *which part* of a model changes when they take place, since this may guide any corrective solutions.

The contribution of this chapter is to take a well-known empirical example (the effect of the 1984 Banking Act on narrow money demand in the UK) and examine it in the light of a new framework for understanding breaks in forecast models, introduced by Castle, Fawcett and Hendry (2010). Within the literature (see, for instance, Hendry and Ericsson (1991)) there is a general understanding of when the Act came into force, and the way in which it altered people's motives to hold money; thus the interest here is

in the adjustment process that followed the break. At the forefront is the idea that in the immediate aftermath of the legislation coming into force, there was only incomplete adjustment in aggregate behaviour, as the full implications of the change were not immediately apparent. Coupled with this, an econometrician trying to model this behaviour at the time would have realised relatively quickly – within three quarters – that *some kind* of change had taken place, but in the absence of a crystal ball, would not have been able to anticipate its effects when building a forecast model. Thus an interesting exercise to undertake is to establish how soon after the break, an econometrician might have been able to learn about the adjustment taking place in the economy, and as a subsidiary issue, how quickly would this learning have improved forecast performance.

6.1 Predictability and information reconsidered

Conventional forecast models are concerned with the conditional expectation of y_{T+1}:

$$\widehat{y}_{T+1|T} = \mathsf{E}[y_{T+1}|\mathbf{x}_T] = \mathsf{f}_{T+1}(\mathcal{I}_T). \tag{6.1}$$

The rationale for using this was discussed earlier, since in a linear model the forecast that minimises the mean-square forecast error is the conditional expectation. However, (6.1) is only valid as a forecast model *if* the functional form of $\mathsf{f}_{T+1}(\cdot)$ is known; in the absence of this knowledge, then it also needs to be forecast. This opens an avenue to augmenting the information set on which the function in (6.1) is conditioned, so that it contains not only information generated by the past history of the independent variables, $\mathcal{I}$, but also a separate information set, denoted $\mathcal{K}$ which contains elements that cause the function $\mathsf{f}(\cdot)$ to shift. The nature of these shifts need not be 'economic' factors in a conventional sense: for example, they could include acts of terrorism, technological changes (such as the emergence of the internet, or the 'Big Bang' episode in the London Stock Exchange), political regime changes and new legislation. Thus the type of information that might enter into $\mathcal{K}$ is feasibly much broader that that entering into $\mathcal{I}$.

To see how this might apply to the scalar linear model above, Castle *et al.* (2010) express $\boldsymbol{\beta}_t$ as:

$$\boldsymbol{\beta}_t = \boldsymbol{\beta}_0 + \boldsymbol{\Gamma}\mathbf{w}_t + \nu_t \quad \text{where } \nu_t \sim \mathsf{IN}\left(0, \sigma_\nu^2\right)$$

so that

$$\mathsf{E}_t[\boldsymbol{\beta}_t|\mathbf{w}_t] = \boldsymbol{\beta}_0 + \boldsymbol{\Gamma}\mathbf{w}_t.$$

In the empirical example discussed below, $\mathbf{w}_t$ is modelled as a step function, which is zero for all periods prior to the Banking Act of 1984, and then jumps to one. In general, though, it need not be an indicator function as it is here. Conditioning the expectation of $\boldsymbol{\beta}_t$ on $\mathbf{w}_t$ highlights the role of the $\mathcal{K}_t$ information set. In this case, $\mathcal{K}_t = \sigma(\mathbf{W}_t)$, which is the sigma-field generated by the history of the $\mathbf{w}_t$ function ($\mathbf{W}_t = \mathbf{w}_{t-1}, \ldots, \mathbf{w}_0$). Considering the conditional expectation

$$\mathsf{E}_t[y_t|\mathbf{x}_{t-1}, \mathbf{w}_t] = \boldsymbol{\beta}_0'\mathbf{x}_{t-1} + \boldsymbol{\Gamma}\mathbf{w}_t\mathbf{x}_{t-1} \tag{6.2}$$

then the role of $\mathbf{w}_t$ in generating shifts in y_t (and thereby inducing forecast failure) is clear: as long as $\mathbf{w}_t = 0$ then the model collapses to a conventional constant parameter regression, with additional input from $\boldsymbol{\Gamma}\mathbf{x}_{t-1}$ entering when $\mathbf{w}_t$ moves to one.

New variables can also become relevant in the aftermath of a break, in addition to existing parameters changing. This can be accommodated in the formulation above by assuming that $\mathbf{x}_t = (\mathbf{z}_t' : \mathbf{v}_t')'$ where $\mathbf{z}_{t-1}$ comprises variables that are relevant throughout (or at least in the pre-break period) and $\mathbf{v}_t$ comprises variables that become relevant, post break. In this case partition $\boldsymbol{\beta}_t$ can be partitioned so that:

$$\boldsymbol{\beta}_t = \begin{pmatrix} \boldsymbol{\beta}_0 \\ \mathbf{0} \end{pmatrix} \mathbf{x}_{t-1} + \boldsymbol{\Gamma}^*\mathbf{w}_t\mathbf{x}_{t-1} + \nu_{t-1}$$

where $\boldsymbol{\Gamma}^* = (\boldsymbol{\Gamma}_z' : \boldsymbol{\Gamma}_v')'$, leaving:

$$\mathsf{E}_t[y_t|\mathbf{x}_{t-1}, \mathbf{w}_t] = \boldsymbol{\beta}_0\mathbf{z}_{t-1} + \boldsymbol{\Gamma}_z\mathbf{w}_t\mathbf{z}_{t-1} + \boldsymbol{\Gamma}_v\mathbf{w}_t\mathbf{v}_{t-1}.$$

This is useful when studying examples such as the money demand model below, where a new variable – the interest rate on sight deposits, R_S – jumps up from zero following the Banking Act, whilst other coefficients on existing variables (in $\mathbf{z}_{t-1}$) remain unchanged. This situation would imply that $\mathbf{\Gamma}_z = 0$, $v_t = R_{S,t}$ and Γ_v is its coefficient in a regression.

Although breaks themselves may be unpredictable with respect to the information set generated by the history of $\mathbf{z}$, in the aftermath of structural change the new information contained in the sigma-fields generated by $\mathbf{W}_t$ and $\mathbf{X}_{t-1}$, will be instrumental in building new forecast models that are robust to the break, in the sense that they do not suffer from continued forecast failure.

As a final point, it is helpful to emphasise the time dating of information sets that are used in conditioning. In general, the expression $\mathsf{f}_T(\mathcal{I}_{T-1})$ from (6.1) demonstrates its dependence on $\mathcal{K}_T$:[1]

$$\mathsf{f}_T(\mathcal{I}_{T-1}, \mathcal{K}_T) = \mathsf{f}_0(\mathcal{I}_{T-1}) + \mathsf{f}_1(\mathcal{I}_{T-1}, \mathcal{K}_T).$$

In this expression, the 'conventional' role of lagged information comes through $\mathsf{f}_0(\mathcal{I}_{T-1})$, which might correspond to $\boldsymbol{\beta}_t'\mathbf{x}_{t-1}$ in (6.2), whilst the role of new information is clear in $\mathsf{f}_1(\mathcal{I}_{T-1}, \mathcal{K}_T)$ which corresponds to $\mathbf{\Gamma}\mathbf{w}_t\mathbf{x}_{t-1}$. Two points are of interest here. First, $\mathcal{I}_{T-1}$ may still be relevant in $\mathsf{f}_1(\cdot)$, whether through pre-break variables (the $\mathbf{z}_{t-1}$ above) or through new variables that are incorporated into an expanded information set $\mathcal{I}$ after the break.[2]

Secondly, although $\mathsf{f}_T(\cdot)$ is a function of lagged $\mathcal{I}$, it is dependent on *contemporaneous* $\mathcal{K}$: hence when a forecast is made of $\mathsf{f}_T(\mathcal{I}_{T-1})$ as in (6.1),

[1] This draws on Castle *et al.* (2010).

[2] An alternative way of expressing this could be to say $\mathcal{I}_{T-1} = \sigma(\mathbf{Z}_{T-1})$, $\mathcal{J}_{T-1} = \sigma(\mathbf{V}_{T-1})$ after the break, and then write:

$$\mathsf{f}_T(\mathcal{I}_{T-1}, \mathcal{J}_{T-1}, \mathcal{K}_T) = \mathsf{f}_0(\mathcal{I}_{T-1}) + \mathsf{f}_1(\mathcal{I}_{T-1}, \mathcal{J}_{T-1}, \mathcal{K}_T).$$

in practice this means:

$$\begin{aligned}\widehat{y}_{T|T-1} &= \mathsf{E}_T[\mathsf{f}_T(\mathcal{I}_{T-1},\mathcal{K}_T)|\mathcal{I}_{T-1},\mathcal{K}_{T-1}] \\ &= \mathsf{f}_0(\mathcal{I}_{T-1}) + \mathsf{E}_T[\mathsf{f}_1(\mathcal{I}_{T-1},\mathcal{K}_T)|\mathcal{K}_{T-1}].\end{aligned}$$

Before a break, the second term on the right-hand side is zero, but the error:

$$\mathsf{E}_T[\mathsf{f}_T(\mathcal{I}_{T-1},\mathcal{K}_T)|\mathcal{I}_{T-1},\mathcal{K}_{T-1}] - \mathsf{E}_T[\mathsf{f}_T(\mathcal{I}_{T-1},\mathcal{K}_T)|\mathcal{I}_{T-1},\mathcal{K}_T]$$

provides valuable information about the nature of any structural change that might have taken place, once a break has happened.

6.2 Forecasting during a break

Consider a DGP given by:[3]

$$y_t = \alpha + \exp\left(\lambda - \psi\left[t - T\right]\right) 1_{\{t \geq T\}} + \epsilon_t \quad \text{with} \quad \epsilon_t \sim \mathsf{IN}\left(0, \sigma_\epsilon^2\right) \tag{6.3}$$

where $1_{\{t \geq T\}}$ is an indicator function which is zero before time T, and unity thereafter, and $\lambda > 0,\ \psi > 0$. The existence of a break at time T is known, but the form it takes is not known, to match section 6.3. The outcome at $T+1$ from (6.3) is:

$$y_{T+1} = \alpha + \exp\left(\lambda - \psi\right) + \epsilon_{T+1}. \tag{6.4}$$

Four alternative 1-step ahead forecasting devices are considered here: (a) an intercept-corrected model; (b) a differenced device; (c) an estimated version of (6.3); and (d) ignoring the break.

6.2.1 Forecasts at $T+1$

6.2.1.1 Intercept correction

Here:

$$y_t = \gamma + \delta D_{\{T\}} + v_t \tag{6.5}$$

[3]This section draws on Castle *et al.* (2010).

where $D_{\{T\}}$ is a dummy variable with the value unity from T onwards, and is zero otherwise. As $D_{\{T\}} = 1$, using the full-sample data, $\widetilde{\gamma}$ is the sample mean over $1 \ldots T-1$, so $\mathsf{E}[\widetilde{\gamma}] = \alpha$:

$$\mathsf{V}[\widetilde{\gamma}] = \frac{\sigma_\epsilon^2}{T-1}. \tag{6.6}$$

and:

$$\widetilde{\delta} = y_T - \widetilde{\gamma} = \alpha - \widetilde{\gamma} + \exp(\lambda) + \epsilon_T \tag{6.7}$$

with $\mathsf{E}\left[\widetilde{\delta}\right] = \exp(\lambda)$ and:

$$\mathsf{V}\left[\widetilde{\delta}\right] = \mathsf{V}[\alpha - \widetilde{\gamma}] + \sigma_\epsilon^2 = \sigma_\epsilon^2 \left(1 + (T-1)^{-1}\right). \tag{6.8}$$

The forecast is $\widetilde{y}_{T+1|T} = \widetilde{\gamma} + \widetilde{\delta}$, so from (6.4):

$$\widetilde{\epsilon}_{T+1|T} = y_{T+1} - \widetilde{y}_{T+1|T} = (\alpha - \widetilde{\gamma}) + \exp(\lambda - \psi) - \widetilde{\delta} + \epsilon_{T+1},$$

with $\mathsf{E}\left[\widetilde{\epsilon}_{T+1|T}\right] = \exp(\lambda)\{\exp(-\psi) - 1\}$ and hence:

$$\begin{aligned} \mathsf{E}\left[\left(\widetilde{\epsilon}_{T+1|T} - \mathsf{E}\left[\widetilde{\epsilon}_{T+1|T}\right]\right)^2\right] &= \mathsf{E}\left[\left((\alpha - \widetilde{\gamma}) + \left[\exp(\lambda) - \widetilde{\delta}\right] + \epsilon_{T+1}\right)^2\right] \\ &= 2\sigma_\epsilon^2 \left(1 + (T-1)^{-1}\right), \end{aligned} \tag{6.9}$$

using (6.8) and (6.6). As found in Chapter 5, this doubles the forecast error variance. Thus, the mean-square forecast error (MSFE) is:

$$\mathsf{M}\left[\widetilde{\epsilon}_{T+1|T}\right] = \exp(2\lambda)\{\exp(-\psi) - 1\}^2 + 2\sigma_\epsilon^2 \left(1 + (T-1)^{-1}\right). \tag{6.10}$$

6.2.1.2 Differenced device

The differenced model is simply:

$$\Delta \overline{y}_{T+1|T} = 0 \tag{6.11}$$

or $\overline{y}_{T+1|T} = y_T$, so no parameter estimation is necessary. From (6.4), letting $\overline{\epsilon}_{T+1|T} = y_{T+1} - \overline{y}_{T+1|T}$:

$$\overline{\epsilon}_{T+1|T} = \exp(\lambda)\{\exp(-\psi) - 1\} + \Delta\epsilon_{T+1} \tag{6.12}$$

with $\mathsf{E}\left[\overline{\epsilon}_{T+1|T}\right] = \exp(\lambda)\{\exp(-\psi) - 1\}$ and:

$$\mathsf{E}\left[\left(\overline{\epsilon}_{T+1|T} - \mathsf{E}\left[\overline{\epsilon}_{T+1|T}\right]\right)^2\right] = \mathsf{E}\left[\left(\epsilon_{T+1} - \epsilon_T\right)^2\right] = 2\sigma_\epsilon^2.$$

Differencing also doubles the forecast error variance, so the MSFE is:

$$\mathsf{M}\left[\overline{\epsilon}_{T+1|T}\right] = \exp(2\lambda)\{\exp(-\psi) - 1\}^2 + 2\sigma_\epsilon^2, \tag{6.13}$$

which dominates the intercept-corrected model.

6.2.1.3 Estimated DGP

Next, a more active approach is to try and model the break process explicitly, by estimating the parameters in the ogive adjustment function. Using the sample mean over $t = 1, \ldots, T-1$ for $\widehat{\alpha}$ in (6.3), this forecast model is:

$$\widehat{y}_{T+1|T} = \widehat{\alpha} + \exp\left(\widehat{\lambda}\right) \tag{6.14}$$

where $\exp(\widehat{\lambda}) = y_T - \widehat{\alpha}$, since ψ cannot be estimated at T. Then $\mathsf{E}[\exp(\widehat{\lambda})] = \exp(\lambda)$, so:

$$\mathsf{E}\left[\exp\left(\widehat{\lambda}\right)\right] = \mathsf{E}\left[\alpha - \widehat{\alpha}\right] + \exp(\lambda) = \exp(\lambda),$$

and:

$$\mathsf{V}\left[\exp\left(\widehat{\lambda}\right)\right] = \mathsf{E}\left[\left((\alpha - \widehat{\alpha}) + \epsilon_T\right)^2\right] = \sigma_\epsilon^2\left(1 + \frac{1}{T-1}\right). \tag{6.15}$$

The forecast error $\widehat{\epsilon}_{T+1|T} = y_{T+1} - \widehat{y}_{T+1|T}$ is:

$$\widehat{\epsilon}_{T+1|T} = (\alpha - \widehat{\alpha}) + \exp(\lambda - \psi) - \exp\left(\widehat{\lambda}\right) + \epsilon_{T+1}$$

so $\mathsf{E}\left[\widehat{\epsilon}_{T+1|T}\right] = \exp(\lambda)\{\exp(-\psi) - 1\}$ and as $\mathsf{V}\left[\widehat{\alpha}\right] = \mathsf{V}\left[\widetilde{\gamma}\right]$, using (6.15) and (6.6):

$$\begin{aligned}\mathsf{E}\left[\left(\widehat{\epsilon}_{T+1|T} - \mathsf{E}\left[\widehat{\epsilon}_{T+1|T}\right]\right)^2\right] &= \mathsf{E}\left[\left((\alpha - \widehat{\alpha}) + \left[\exp(\lambda) - \exp\left(\widehat{\lambda}\right)\right] + \epsilon_{T+1}\right)^2\right] \\ &= 2\sigma_\epsilon^2\left(1 + (T-1)^{-1}\right). \end{aligned} \tag{6.16}$$

The MSFE is:

$$\mathsf{M}\left[\widehat{\epsilon}_{T+1|T}\right] = \exp(2\lambda)\left\{\exp(-\psi) - 1\right\}^2 + 2\sigma_\epsilon^2\left(1 + (T-1)^{-1}\right). \quad (6.17)$$

Surprisingly, (6.10) and (6.17) are identical, and (6.13) is only smaller by $2\sigma_\epsilon^2/(T-1)$, a small 'saving' by not estimating any parameters. Thus, the forecast-error bias is identical for all three methods, and is zero only if $\psi = 0$, so the only difference is from estimation uncertainty.

6.2.1.4 An unadjusted model

Finally:

$$\overleftarrow{y}_{T+1|T} = \overleftarrow{\alpha} \quad (6.18)$$

where $\overleftarrow{\alpha} = \sum_{t=1}^{T} y_t/T$, then:

$$\mathsf{E}[\overleftarrow{\alpha}] = \alpha + \frac{1}{T}\exp(\lambda) \quad \text{and} \quad \mathsf{V}[\overleftarrow{\alpha}] = \frac{\sigma_\epsilon^2}{T},$$

so $\overleftarrow{\epsilon}_{T+1|T} = y_{T+1} - \overleftarrow{y}_{T+1|T}$ with:

$$\mathsf{E}\left[\overleftarrow{\epsilon}_{T+1|T}\right] = \exp(\lambda - \psi) - \frac{1}{T}\exp(\lambda)$$

and MSFE:

$$\mathsf{M}\left[\overleftarrow{\epsilon}_{T+1|T}\right] = \exp(2\lambda)\left(\exp(-\psi) - \frac{1}{T}\right)^2 + \sigma_\epsilon^2\left(1 + \frac{1}{T}\right) \quad (6.19)$$

The ranking varies with the DGP parameter values, specifically the size of the break relative to σ_ϵ^2, but in general (6.19) will be larger than the first three approaches, which underlines the insurance benefit of adjusting to a break.

6.2.2 Forecasting at $T+2$

One period later, the DGP is now:

$$y_{T+2} = \alpha + \exp(\lambda - 2\psi) + \epsilon_{T+2}. \quad (6.20)$$

Models (a) to (c) continue to adjust following the break, so it is useful to examine their forecast errors.

6.2.2.1 Intercept correction

Updating using (6.5), $\widetilde{\gamma}$ remains as before, but:

$$\widetilde{\delta} = \frac{1}{2}(y_T + y_{T+1}) - \widetilde{\gamma} = \alpha - \widetilde{\gamma} + \frac{1}{2}(\exp(\lambda) + \exp(\lambda - \psi)) + \frac{1}{2}(\epsilon_T + \epsilon_{T+1}) \tag{6.21}$$

so $\mathsf{E}[\widetilde{\delta}] = \exp(\lambda)(1 + \exp(-\psi))/2$ and:

$$\begin{aligned}\mathsf{V}\left[\widetilde{\delta}\right] &= \mathsf{V}[\alpha - \widetilde{\gamma}] + \frac{\sigma_\epsilon^2}{2} \\ &= \sigma_\epsilon^2\left(\frac{1}{2} + (T-1)^{-1}\right).\end{aligned}$$

The forecast is $\widetilde{y}_{T+2|T+1} = \widetilde{\gamma} + \widetilde{\delta}$, with error $\widetilde{\epsilon}_{T+2|T+1} = y_{T+2} - \widetilde{y}_{T+2|T+1}$ so:

$$\widetilde{\epsilon}_{T+2|T+1} = (\alpha - \widetilde{\gamma}) + \exp(\lambda - 2\psi) - \widetilde{\delta} + \epsilon_{T+2}$$

where $\mathsf{E}\left[\widetilde{\epsilon}_{T+2|T+1}\right] = \exp(\lambda)\left[\exp(-2\psi) - \frac{1}{2}(1 + \exp(-\psi))\right]$, and hence:

$$\begin{aligned}\mathsf{E}\left[\left(\widetilde{\epsilon}_{T+2|T+1} - \mathsf{E}\left[\widetilde{\epsilon}_{T+2|T+1}\right]\right)^2\right] &= \frac{1}{2}\sigma_\epsilon^2 + \sigma_\epsilon^2(T-1)^{-1} + \sigma_\epsilon^2 \\ &= \sigma_\epsilon^2\left(\frac{3}{2} + (T-1)^{-1}\right),\end{aligned}$$

which is lower than before, since each lagged DGP error enters with a weight of 1/2. Thus the MSFE is:

$$\begin{aligned}\mathsf{M}\left[\widetilde{\epsilon}_{T+2|T+1}\right] = \exp(2\lambda)&\left[\exp(-2\psi) - \frac{1}{2}(1 + \exp(-\psi))\right]^2 \\ &+ \sigma_\epsilon^2\left(\frac{3}{2} + (T-1)^{-1}\right).\end{aligned} \tag{6.22}$$

6.2.2.2 Differencing

The differenced model remains:

$$\overline{y}_{T+2|T+1} = y_{T+1} \tag{6.23}$$

so letting $\overline{\epsilon}_{T+2|T+1} = y_{T+2} - \overline{y}_{T+2|T+1}$:

$$\overline{\epsilon}_{T+2|T+1} = \exp(\lambda)\left[\exp(-2\psi) - \exp(-\psi)\right] + \Delta\epsilon_{T+1} \tag{6.24}$$

with $\mathsf{E}\left[\overline{\epsilon}_{T+2|T+1}\right] = \exp(\lambda)\left[\exp(-2\psi) - \exp(-\psi)\right]$ and:

$$\mathsf{E}[\left(\overline{\epsilon}_{T+2|T+1} - \mathsf{E}\left[\overline{\epsilon}_{T+2|T+1}\right]\right)^2] = 2\sigma_\epsilon^2.$$

This is now larger than the variance of the intercept-corrected model, although the MSFE is:

$$\mathsf{M}\left[\overline{\epsilon}_{T+2|T+1}\right] = \exp(2\lambda)\left[\exp(-2\psi) - \exp(-\psi)\right]^2 + 2\sigma_\epsilon^2. \tag{6.25}$$

As $\psi > 0$, differencing generates smaller bias than using the intercept-corrected model, so the innovation variance σ_ϵ^2 determines whether (6.25) dominates (6.22) at $T+2$.

6.2.2.3 Estimated DGP

Now ψ can be estimated, albeit with considerable uncertainty, so:

$$\widehat{y}_{T+2|T+1} = \widehat{\alpha} + \exp\left(\widehat{\lambda}\right)\exp\left(-\widehat{\psi}\right) \tag{6.26}$$

where $\widehat{\alpha}$, $\exp(\widehat{\lambda}) = y_T - \widehat{\alpha}$, and $\mathsf{V}[\exp(\widehat{\lambda})] = \sigma_\epsilon^2(1 + (T-1)^{-1})$ remain as before, with:

$$\exp\left(-\widehat{\psi}\right) = \left(y_{T+1} - \widehat{\alpha}\right) / \exp\left(\widehat{\lambda}\right).$$

Thus, to a first approximation:

$$\begin{aligned}\mathsf{E}\left[\exp\left(-\widehat{\psi}\right)\right] &= \mathsf{E}\left[\exp\left(-\widehat{\lambda}\right)\left(\alpha - \widehat{\alpha} + \exp(\lambda - \psi) + \epsilon_{T+1}\right)\right] \\ &\approx \exp(-\psi),\end{aligned}$$

and:

$$\begin{aligned}\mathsf{V}\left[\exp\left(-\widehat{\psi}\right)\right] &= \mathsf{E}\left[\left(\exp\left(-\widehat{\psi}\right) - \mathsf{E}\left[\exp\left(-\widehat{\psi}\right)\right]\right)^2\right] \\ &\approx \exp(-2\lambda)\,\sigma_\epsilon^2\left(1 + \frac{1}{T-1}\right).\end{aligned} \tag{6.27}$$

The forecast error $\widehat{\epsilon}_{T+2|T+1} = y_{T+2} - \widehat{y}_{T+2|T+1}$ is:

$$\widehat{\epsilon}_{T+2|T+1} = (\alpha - \widehat{\alpha}) + \exp(\lambda - 2\psi) - \exp\left(\widehat{\lambda} - \widehat{\psi}\right) + \epsilon_{T+2},$$

so $\mathsf{E}\left[\widehat{\epsilon}_{T+2|T+1}\right] = \exp(\lambda)\left[\exp(-2\psi) - \exp(-\psi)\right]$ and neglecting the covariance between $\widehat{\lambda}$ and $\widehat{\psi}$:

$$\begin{aligned}\mathsf{E}\left[\left(\widehat{\epsilon}_{T+2|T+1} - \mathsf{E}\left[\widehat{\epsilon}_{T+2|T+1}\right]\right)^2\right] &= \mathsf{E}\left[(\alpha - \widehat{\alpha} + \exp(\lambda)\exp(-\psi)\right. \\ &\quad \left. - \exp\left(\widehat{\lambda}\right)\exp\left(-\widehat{\psi}\right) + \epsilon_{T+2}\right)^2\Big] \\ &\approx \sigma_\epsilon^2\left(1 + (T-1)^{-1}\right)(2 + \exp(-2\lambda))\,. \end{aligned} \tag{6.28}$$

Thus, the MSFE is:

$$\begin{aligned}\mathsf{M}\left[\widehat{\epsilon}_{T+2|T+1}\right] &= \exp(2\lambda)\left[\exp(-2\psi) - \exp(-\psi)\right]^2 \\ &\quad + \sigma_\epsilon^2\left(1 + (T-1)^{-1}\right)(2 + \exp(-2\lambda))\,. \end{aligned} \tag{6.29}$$

This will exceed $\mathsf{M}\left[\overline{\epsilon}_{T+2|T+1}\right]$ due to the cost of estimating the additional parameter ψ, so even after two periods and for an 'ogive' break, the robust methods will still outperform.

6.3 Learning and forecasting: an empirical example

The phenomena discussed above – in particular the distinction between information sets – are illustrated in an application to data on money demand in the UK. This revolves around the Banking Act of 1984, which altered the opportunity cost of holding money, since it simultaneously allowed banks to pay interest on checking accounts, but required them to report interest income payments to the Inland Revenue. The latter change affected individuals who had hitherto not paid income tax on interest income (previously banks were not obliged to report). By switching wealth to checking accounts, though, they were able to earn R_S interest on deposits, whilst attributing earlier non-payment of tax on interest income to the fact that banks had been prevented from paying any interest on such accounts in the first place. In combination, these forces provided both the demand and supply for a shift in assets from non-M1 deposits to M1.

The break caused by the Banking Act is relevant to the framework de-

veloped above and in Chapter 3, in two respects. First, the dramatic shift in the opportunity cost is a significant non-linearity that needs to be modeled correctly to avoid systematic forecast failure. Secondly, since the cause of the change was legislative, and knowledge of it did not spread immediately across the economy, the case offers potential insights into modeling forces that induce structural change through the $\mathcal{K}_T$ information set.

This raises three questions:

1. How quickly did agents respond to the shift in the opportunity cost, and at what point could an econometrician have been able to learn about this?

2. Provided that they succeeded in modeling the adjustment process, how soon could an econometrician have used this updated knowledge to forecast better; and

3. Regardless of any need to learn about a shift, could a 'mechanistic' transformation of adapted model have produced better forecasts than the unadjusted model?

To answer the first, a weighting function is used, to shed light on the adjustment process in the aftermath of the break. The second embeds this understanding of learning adjustment into a forecast model for real money and compares the resulting predictions to the alternative of no updating or adjustment. Finally, the 'mechanistic' correction mechanisms discussed elsewhere in this thesis, and particularly in Chapter 4 are explained, addressing the third question.

6.3.1 A model of money demand

The underlying model of money demand follows that of Hendry and Mizon (1993) and Hendry and Doornik (1994) (also see Hendry (1979), Hendry and Ericsson (1991), Boswijk (1992), Boswijk and Doornik (2004) and Hendry (2006)), and uses a cointegrated system to model the following variables:

M	nominal M1
Y	real total final expenditure (TFE) at 1985 prices
P	the TFE deflator
R_{LA}	the three-month local authority interest rate
R_{S}	the interest rate on sight bank deposits

The theoretical basis is a model which links demand for real money, $\mathsf{m}-\mathsf{p}$ (lowercase denoting logs) to (log) income y (transactions motive) and the interest rate R_{LA} on an 'outside option' (i.e. measuring the opportunity cost of holding money).

In the run-up to the Banking Act (from 1964(3) to 1984(3)), the estimated long-run relationship $\boldsymbol{\beta}'\mathbf{x}_t$, where $\boldsymbol{\beta}$ is the matrix of cointegrating vectors, and $\mathbf{x}_t = (\mathsf{m}-\mathsf{p}_t, \mathsf{y}_t, \Delta\mathsf{p}_t, \mathsf{R}_{LA,t})'$ is given by:[4]

$$\boldsymbol{\beta}'\mathbf{x}_t = \begin{pmatrix} \underset{(-)}{1.00}(\mathsf{m}-\mathsf{p})_t - \underset{(-)}{1.00}\mathsf{y}_t + \underset{(1.54)}{6.32}\Delta\mathsf{p}_t + \underset{(0.65)}{6.97}\mathsf{R}_{LA,t} \\ \underset{(-)}{1.00}\mathsf{y}_t + \underset{(0.33)}{1.25}\mathsf{R}_{LA,t} - \underset{(-)}{0.0062}t \end{pmatrix}. \tag{6.30}$$

The upper row of $\boldsymbol{\beta}'\mathbf{x}_t$ represents an 'excess money demand' relationship, showing the demand for real money balances in excess of the amount required to pay for goods and services, whilst the lower row can be interpreted as a trend output relationship, showing excess demand for real goods and services. Unique identification of the system is achieved by imposing the same restrictions on parameters as Hendry (1995, Chapter 16) (*inter alia*).

6.3.2 Structural breaks and agent learning

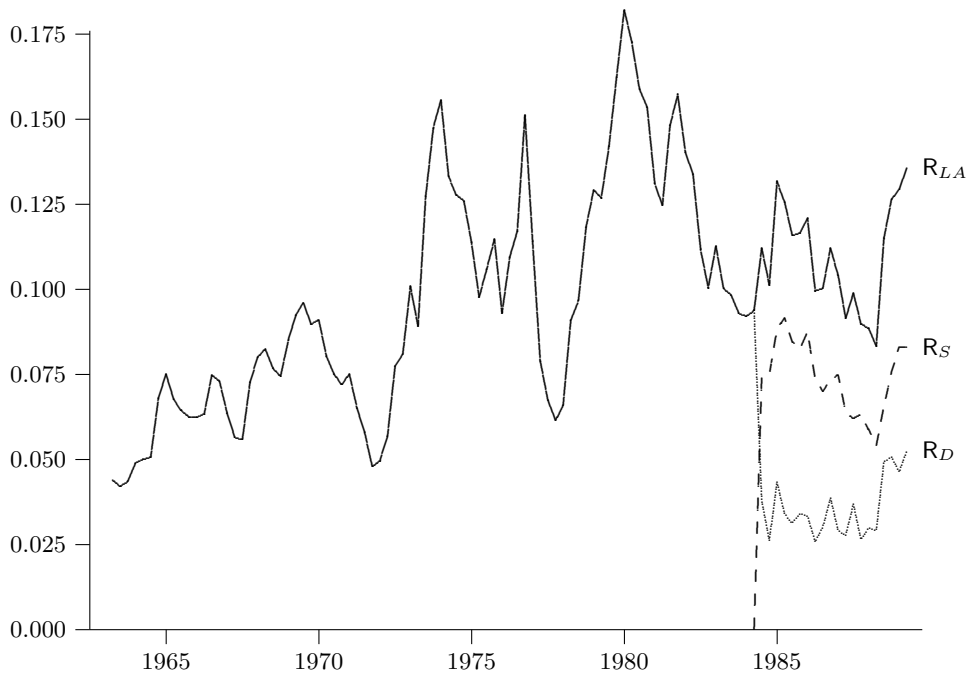
The impact of the Banking Act on the opportunity cost of holding money is demonstrated in Figure 6.1, which compares R_{LA} to R_{S}, and calculates the differential

$$\mathsf{R}_{D} = \mathsf{R}_{LA} - \mathsf{R}_{S}. \tag{6.31}$$

This experienced a structural break, since as soon as banks were permitted to pay interest on checking accounts, they did so (in the face of competition from overseas banks, who were already paying interest). This drove a 3%

[4] This includes restrictions on the coefficient on y in the upper row, and on $\mathsf{m}-\mathsf{p}$, y and the trend in the lower row. These are accepted in a χ^2 test with p-value 0.4681.

Figure 6.1: OPPORTUNITY COST OF HOLDING MONEY: COMPARISONS



NOTES: R_{LA} is the interest rate on deposits with local authorities; R_S is the interest rate on sterling retail sight deposits, and $R_D = R_{LA} - R_S$, the interest rate differential between the two.

wedge between R_{LA} and R_S. A pertinent question here is whether agents in the economy adjusted immediately to the change, or whether they took time to learn about it.

In order to find an answer, a suitable mechanism to capture agent learning is a smooth transition function of the kind studied in Section 6.2, in which a weight w_t is given by

$$w_t = \begin{cases} (1 + \exp[\alpha - \beta(t - t^* + 1)])^{-1} & \text{for } t \geq t^* = 1984(3) \\ 0 & \text{otherwise} \end{cases}$$

This is then used to construct a 'net' interest rate R_N, where

$$R_N = R_{LA} - w_t R_S. \tag{6.32}$$

Thus as time elapses, for $\alpha, \beta > 0$, w_t tends towards one, and R_N moves towards R_D, with the speed of adjustment determined by α and β.

The advantage of this formulation is that it offers a smooth transition function to model changes in the opportunity cost measure, whilst requiring a further two parameters to be estimated in addition to any existing parameters in the model. Further, recursive estimation of α and β should reveal how quickly agents could have learnt of the change in functional form in the money demand function, which is a separate issue to how quickly R_N adjusted to resemble R_D. The difference between the two questions is that the former asks how quickly 'stable' parameter estimates could have been formed, whilst the latter examines how quickly w_t reached one after the Banking Act.

Following Hendry and Ericsson (1991), the parameters α and β are estimated via a single-equation specification of the form:

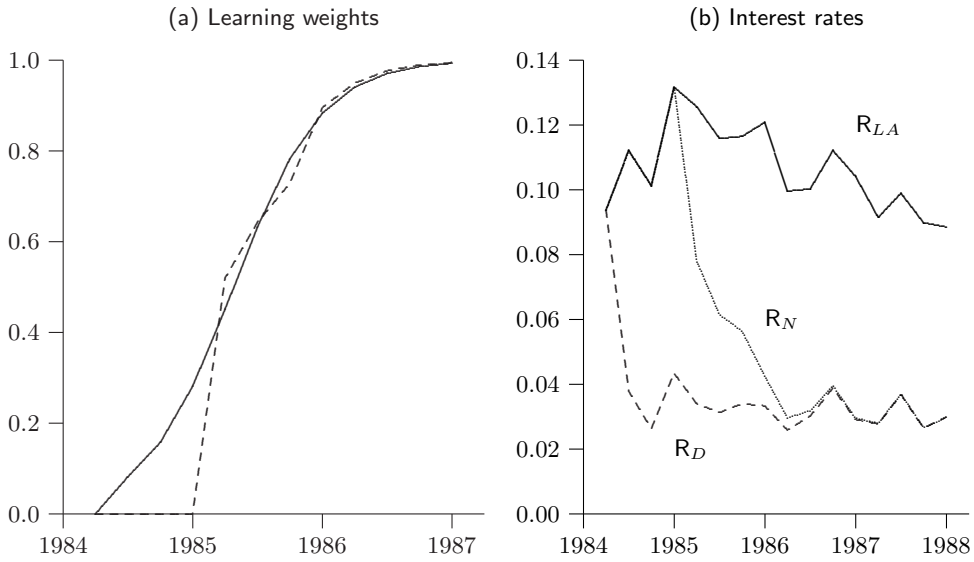
$$\begin{aligned}\Delta(\mathsf{m}-\mathsf{p})_t = \mu_0 &+ \mu_1\Delta(\mathsf{m}-\mathsf{p})_{t-1} + \mu_2\Delta\mathsf{p}_t + \mu_3\Delta\mathsf{y}_{t-1} \\ &+ \mu_4(\mathsf{R}_{LA,t} - (1+\exp(\alpha+\beta(t-t^*+1)))^{-1}\mathsf{R}_{S,t}) \\ &+ \mu_5(\mathsf{m}-\mathsf{p}-\mathsf{y})_{t-2} + \epsilon_t \end{aligned} \tag{6.33}$$

Using the estimates for α and β from this, the 'learning-adjusted' interest rate R_N can then be embedded in the cointegrating system, in place of R_{LA} used above. Over the whole sample period (1964(3) to 1989(2)), (6.33) can be estimated using PcGive (see Doornik and Hendry (2007)) yielding:

$$\begin{aligned}\widehat{\Delta(\mathsf{m}-\mathsf{p})}_t = &\underset{(0.0046)}{0.0252} - \underset{(0.0745)}{0.290}\,\Delta(\mathsf{m}-\mathsf{p})_{t-1} - \underset{(0.130)}{0.696}\,\Delta\mathsf{p}_t + \underset{(0.098)}{0.224}\,\Delta\mathsf{y}_{t-1} \\ &- \underset{(0.068)}{0.657}\left(\mathsf{R}_{LA,t} - \left(1+\exp\left(\underset{(1.667)}{3.16} - \underset{(0.376)}{0.740}(t-t^*+1)\right)\right)^{-1}\mathsf{R}_{S,t}\right) \\ &- \underset{(0.009)}{0.095}(\mathsf{m}-\mathsf{p}-\mathsf{y})_{t-2}. \end{aligned} \tag{6.34}$$

The coefficient estimates for lagged money growth, inflation, output growth and the error correction term are similar to the linear estimation in the previous section, and the estimates of $\widehat{\alpha}$ and $\widehat{\beta}$ at 3.16 and 0.740 respectively are consistent with previous studies in the literature: Hendry and Ericsson (1990, Appendix B) found estimates of 5 and 1.2 for the same coefficients using US data.

Figure 6.2: LEARNING WEIGHTS



NOTE: R_N is the net interest rate given by $R_{LA} - w_t R_S$, using the recursive w_t weights from the left hand panel. R_D is the interest rate differential $R_{LA} - R_S$, measuring the difference between the local authority interest rate, and the rate on sight deposits. The solid line in panel (a) shows the weight constructed with whole-sample estimates of α and β, whilst the dashed line shows an alternative weight series constructed with recursive estimates.

Using these estimates, it is possible to construct a weight series applying the same $\widehat{\alpha}$ and $\widehat{\beta}$ at all points following 1984(3); this is shown in Figure 6.2. In line with the definition of w_t above, the weight is zero before 1984(3), climbs to approximately 0.6 within four quarters of t^*, and is very close to one after eight quarters.

Recursive non-linear estimation is helpful in assessing the stability of $\widehat{\alpha}$ and $\widehat{\beta}$ estimates, and there is evidence of instability for two periods following the break, in 1984(4) and 1985(1). This can be explained by the small number of observations available relative to the number of parameters (two) that need to be estimated.[5] However, beyond this there is remarkable stability in the estimates, and the equation passes Chow tests of parameter constancy.

In contrast to the full-sample weight in Figure 6.2 (a), the dashed line shows a 'recursive' weight series constructed by inserting the recursive $\widehat{\alpha}$

[5]Of course, neither α nor β could be identified before 1984(3).

and $\widehat{\beta}$ into the weight function.[6] The resulting recursive w_t differs from the original weight series only slightly, once the effect of setting the first few post-Banking Act observations to zero is taken into account.

In answer to the first question posed above, it would seem that by 1985(2), agents were in a position to form an accurate impression of the components of the weighting function w_t. However, it was only by early 1986 that they actually adjusted to the interest rate differential, as shown by the weight reaching close to one in Figure 6.2. Panel (b) also demonstrates this learning process, since it reveals a significant gap between R_D and R_N from 1984(3) to 1986(2). Thus it is possible to split the learning process into two phases: the first involving understanding the nature of the structural change, and secondly, using this understanding to adapt to it.

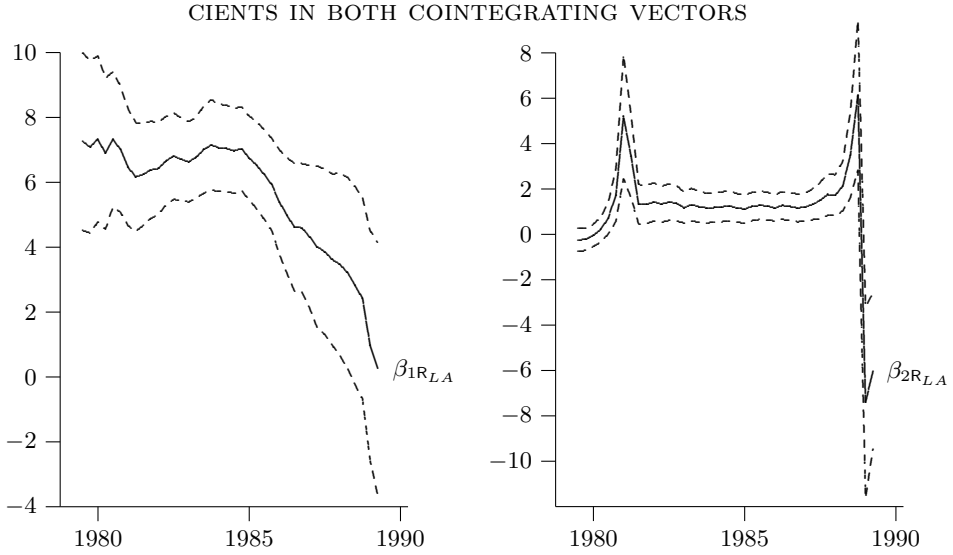
6.3.3 Forecasting after the break

In order to answer the second question, the baseline model above can be compared to a learning-adjusted model as both tried to forecast, after the Banking Act. As in the previous section, a recursive approach lends itself well to this objective, and the first step is to examine the progressive effect of the break on the cointegrating model itself.

Of the unrestricted parameters in (6.30), the two of greatest interest are the interest rate coefficients, denoted $\beta_{1\mathsf{R}_{LA}}$ for that of the upper row, and $\beta_{2\mathsf{R}_{LA}}$ for that in the lower row. The discussion above speculated that the effect of the Act was to alter the *measure* of the opportunity cost of holding money, without changing the actual relationship given by the coefficients in the cointegrating vector. Thus an interesting issue to pursue is whether $\beta_{1\mathsf{R}_{LA}}$ and $\beta_{2\mathsf{R}_{LA}}$ changed over time after 1984(3), and whether the same coefficients in the system comprising R_N instead of R_{LA} exhibited the same behaviour. If the data suggested stability in the latter case that was absent in the former, this would suggest that even during and after the break, the

[6] In order to address the issue of discarding the two observations before 1985(2), w_t is set to zero for each. The rationale for this is that since agents could not obtain 'reasonable' estimates for α and β, it is implausible to think that they used them to update their learning function.

Figure 6.3: RECURSIVE ESTIMATES OF INTEREST RATE COEFFICIENTS IN BOTH COINTEGRATING VECTORS



NOTE: $\beta_{1\mathsf{R}_{LA}}$ refers to the coefficient on R_{LA} in the upper row of $\boldsymbol{\beta}'\mathbf{x}_t$; $\beta_{2\mathsf{R}_{LA}}$ refers to the coefficient in the lower row of the same vector. Dashed lines capture parameter uncertainty by showing upper and lower bounds of two standard errors on each estimate.

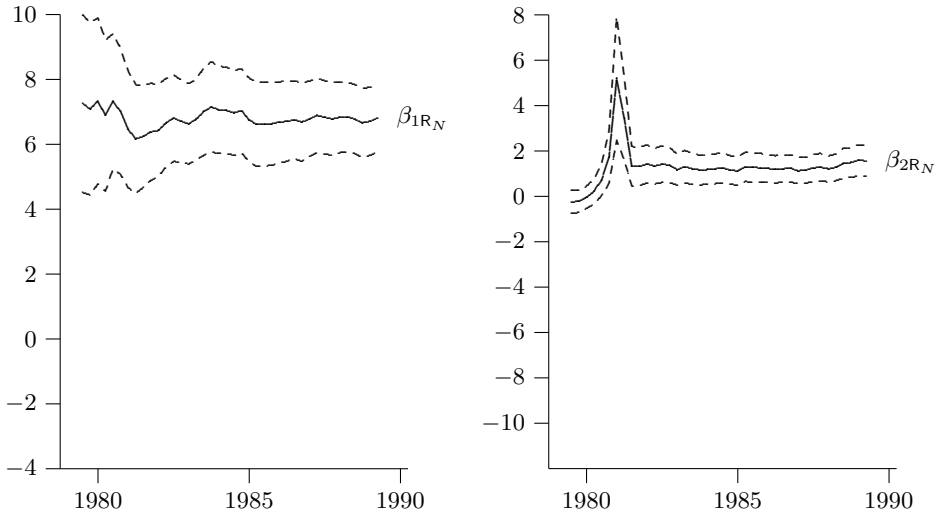
underlying behavioural relationship determining equilibrium in the economy did not change. This in itself is significant because the explanation for the structural break did not lie within the information set $\mathcal{I}_t$ up to 1984(3), but rather came from an exogenous legislative source, raising two issues.

First, the function $f_t(\cdot)$ changed such that $f_{t^*}(\cdot) \neq f_{t^*-k}(\cdot)$ for $k = 1, 2, \ldots$ in such a way that even knowing $\mathcal{I}_{t^*}$ would not be sufficient to model the shift. Rather, information in the separate set $\mathcal{K}_{t^*}$ would be required.

Secondly, the fact that the source of the break could not be found in $\mathcal{I}_t$ would suggest that regime-switching models based on endogenous switching criteria could not adequately capture the change.

Figure 6.3 presents recursive estimates for the interest rate coefficients, which point to significant parameter instability in $\beta_{1\mathsf{R}_{LA}}$ after t^*, and $\beta_{2\mathsf{R}_{LA}}$ at the end of the sample. The marked fall in $\beta_{1\mathsf{R}_{LA}}$ from a range of 6-7 up to 1985 down to nearly zero by 1989 supports the argument that the cointegrating vector as it is, is mis-specified, so that the fall in the opportunity cost of holding money is reflected in the interest rate coefficient.

Figure 6.4: RECURSIVE ESTIMATES OF INTEREST RATE COEFFICIENTS IN BOTH COINTEGRATING VECTORS



NOTES: $\beta_{1\mathsf{R}_N}$ refers to the coefficient on R_N in the upper row of $\boldsymbol{\beta}'\mathbf{x}_t$; $\beta_{2\mathsf{R}_N}$ refers to the coefficient in the lower row of the same vector. Dashed lines capture parameter uncertainty by showing upper and lower bounds of two standard errors on each estimate. The scale of both panels matches those in Figure 6.3.

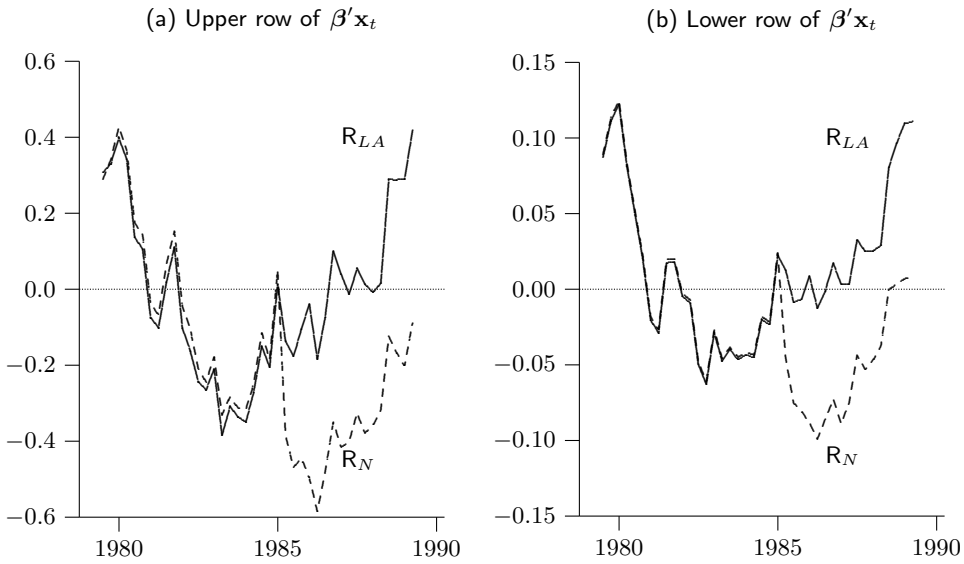
Replacing R_{LA} with R_N in the vector $\mathbf{x}_t$ yields a dramatic improvement in parameter stability, suggesting that with the correct measure in place, the money demand relationship was stable.

What bearing does the foregoing discussion have on the question of agent learning and forecasting? It is relevant to the former, since parameter stability suggests that R_N was the correct measure of the opportunity cost of holding money. Since both R_{LA} and R_S, which comprise two of the three elements of R_N (as equation (6.32) shows), were public knowledge, it was possible in theory for agents to learn about R_N. In practice, the weighting function w_t allows us to quantify how quickly this learning took place.

The relevance to forecasting is clear from Figure 6.5, which contrasts the extent of disequilibrium given by $\boldsymbol{\beta}'\mathbf{x}_t$, when $\mathbf{x}_t$ includes R_N as opposed to R_{LA}. In both panels (a) and (b), which represent the first and second rows respectively of $\boldsymbol{\beta}'\mathbf{x}_t$, the extent (and in (a) even the sign) of disequilibrium differs substantially.

Since the growth in real money responds to disequilibrium in the upper

Figure 6.5: COINTEGRATING VECTOR DISEQUILIBRIUM



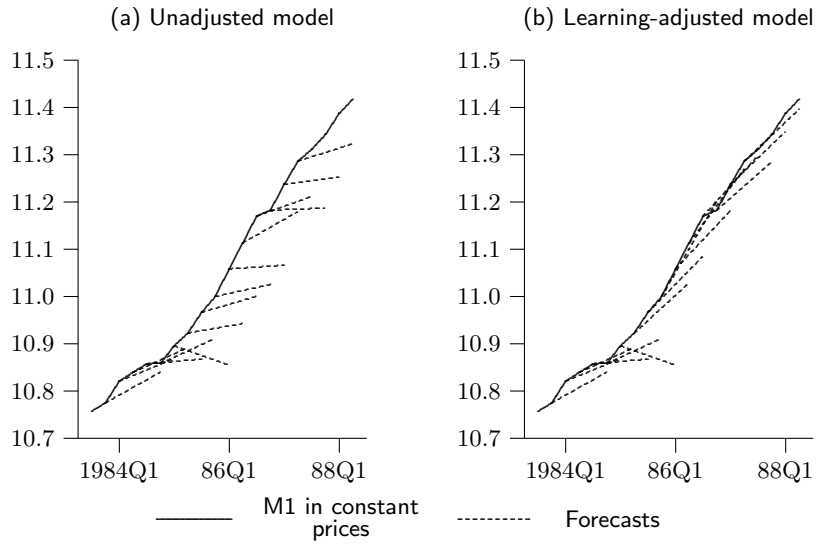
NOTE: The lines denoted R_{LA} represent the elements of the vector $\boldsymbol{\beta}'\mathbf{x}_t$ when R_{LA} is used as the interest rate; R_N refers to the case when the net interest rate is used. In each case, the vectors have been standardised, so they represent deviations from their respective means over the sample period 1964(3) to 1985(4).

row of $\boldsymbol{\beta}'\mathbf{x}_t$, the effect of estimating incorrectly its extent or even sign is severe. As Figure 6.6 shows, forecasts based on a vector equilibrium correction model using R_{LA} underestimated growth in real money since the equilibrium correction element (shown in panel (a) of Figure 6.5) suggested that the system was either at equilibrium, or in a state of positive disequilibrium between 1985 and 1990. In contrast, forecasts based on R_N captured real money growth more accurately since the model suggested that the economy was in negative disequilibrium. This produces the 'hedgehog' forecast spikes in panel (a) of Figure 6.6, which is similar to the forecast failure demonstrated in Figure 3.3. In contrast, the learning-adjusted forecasts suffer from failure at the point of the break, but are rapidly set back on track.

A similar picture is painted with interest rate forecasts, which responded to the lower row of $\boldsymbol{\beta}'\mathbf{x}_t$ shown in panel (b) of Figure 6.5.

To establish *when* agents were able to use their understanding of the

Figure 6.6: REAL MONEY FORECAST COMPARISONS

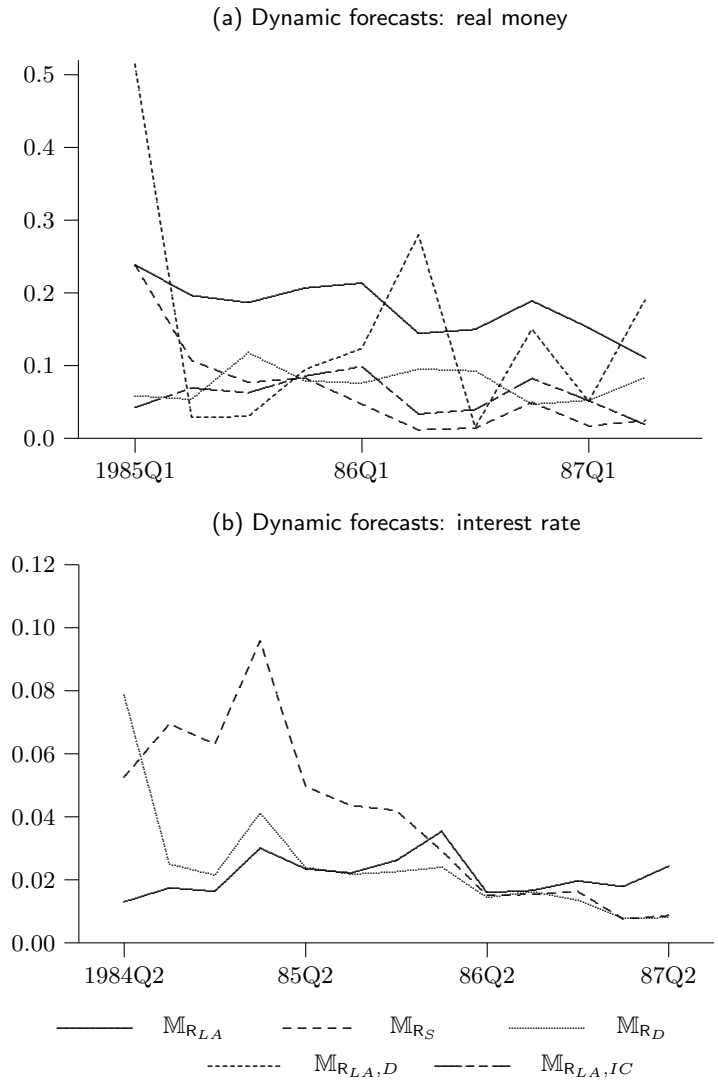


NOTES: Panels show 4-step forecasts from the cointegrated model without break adjustment (in (a)), and with break adjustment (in (b)).

structural change that took place, in order to forecast more accurately the money stock, an illuminating approach is to estimate the model economy recursively and produce a set of forecasts based on the results. The procedure involves estimating the cointegrating relationship over an expanding estimation window starting in 1964(3), and then producing eight multi-step forecasts at each point, recording their root mean squared error. As Figure 6.7 demonstrates, this was done for both real money and interest rate forecasts, comparing the effect of using three different interest rate series in the model: the baseline 'no learning' case (R_{LA}), the progressive learning case (R_N) and the immediate adjustment case (R_D).

From the real money forecasts in panel (a), the first observation to make is that in general, performance is better when the model used takes into account the shift in the opportunity cost of holding money. The dynamic forecasts of the learning-adjusted model $\mathbb{M}_{\mathsf{R}_N}$ are initially identical to those of $\mathbb{M}_{\mathsf{R}_{LA}}$ up to 1985(1), since $w_t = 0$ until the subsequent quarter, but remarkably, *as soon as* adjustment started to take place, forecast performance improved substantially. Even though $w_{85(2)} = 0.519$, indicating only partial

Figure 6.7: FORECAST COMPARISONS



NOTE: Each series represents the root mean squared error (RMSE) of a sequence of forecasts. First the cointegrating relationship within the four variable system is estimated from 1964(3) to one of the points on the graph, such as 1985(2). Then, using the estimated results, eight forecasts are produced and their RMSE is recorded. These are multi-step forecasts, from one to eight quarters ahead. As the recursive estimation window expands from 1984 to 1987, a series of RMSEs is obtained. The subscript on $\mathbb{M}$ in the legend indicates which interest rate series was used in the four variable system. D represents a differenced-VEqCM model, and IC an intercept-corrected model.

adjustment of R_N towards R_D, the RMSE was 45% lower than that of the unadjusted system. After this point there is further improvement in forecast

performance, although not on the same scale, reinforcing the point that only three quarters after the Banking Act, it was feasible that an agent wishing to forecast real money could have done so with greater accuracy by using a learning-adjusted interest rate series.

Secondly, the forecasts from $\mathbb{M}_{\mathsf{R}_D}$ are substantially better than both alternatives from 1984(3) to 1985(1), reflecting the effect of immediate adjustment in the behaviour of at least some holders of money in the economy. An interesting related issue is the fact that the RMSE of $\mathbb{M}_{\mathsf{R}_N}$ is marginally higher than that of $\mathbb{M}_{\mathsf{R}_D}$ for three quarters in 1985. Since this implies that a model assuming immediate adjustment could forecast better than one with a smooth-transition adjustment process, it is worth further comment.

One feature of the dynamic forecasts that in particular to forecast origins from 1985(2) to 1986(1) is that at each origin the adjustment process in the weights w_t is truncated. Thus for example, the eight multistep forecasts produced at 1985(2) all assume that $w_t = 0.519$, so in the econometrician's model, there is no further adjustment of R_N towards R_D. As the origin moves on, w_t increases, so that later forecasts *do* incorporate more complete adjustment.

In order to allow for this effect, a series of one-step forecasts can be compared from 1984(3) to 1987(3). As the forecast origin advances, w_t tends towards one, thereby capturing the adjustment process in R_N as it moves from R_{LA} to R_D. The RMSE for this set of 13 forecasts is lower from $\mathbb{M}_{\mathsf{R}_N}$ than for $\mathbb{M}_{\mathsf{R}_D}$, although the difference is not statistically significant from zero in a Morgan-Granger-Newbold test of comparative forecast accuracy.[7] However, with respect to $\mathbb{M}_{\mathsf{R}_{LA}}$, the $\mathbb{M}_{\mathsf{R}_N}$ forecasts are significantly more accurate using the same test of comparative performance.[8]

The contrast with mechanistic devices such as the differenced VEqCM or intercept-corrected models is shown in panel (a). In the aftermath of the break, the intercept-corrected mechanism performs very well, delivering the

[7]The RMSE for $\mathbb{M}_{\mathsf{R}_N}$ was 0.0204 and for $\mathbb{M}_{\mathsf{R}_D}$, 0.0246.

[8]The Morgan-Granger-Newbold test statistic is -2.76, compared to a critical value of -2.56.

lowest RMSE of all the models considered. Although the DVEqCM forecasts start badly, they are rapidly corrected, demonstrating the usefulness of an automatic device immediately after a break. However, as time progresses, the learning-adjusted model dominates, suggesting a role for direct modeling of the break.

This assessment can be rounded off by discussing the implications of panel (b) in Figure 6.7 which reveals that $\mathbb{M}_{\mathsf{R}_N}$ was poor at forecasting R_N, relative to other models predicting their own interest rates, until 1986. A number of factors contribute to this outcome, and it is worth unpicking each.

First, the effect of comparing dynamic, rather than one-step, forecasts, again has an impact through the adjustment of the weight w_t. Thus at each forecast horizon the econometrician's model assumes that no further adjustment in w_t takes place, whilst in reality w_t tends towards one over the forecast horizon. This leads to $\mathbb{M}_{\mathsf{R}_N}$ failing to model the step shift in the opportunity cost that occurs in 1984(3) until the second quarter of 1985, whereupon there is a significant improvement in dynamic forecast performance.

Thus whilst $\mathbb{M}_{\mathsf{R}_{LA}}$ is more capable of forecasting its opportunity cost measure, the fact remains that this was the *incorrect* one to use, as the parameter instability tests revealed. As a result, the main structural break in the model is missed, leading to the forecast results for real money studied above.

6.4 Concluding comments

In conclusion, this chapter has found that the causes of structural shifts are not restricted to economic factors. Legislative change, in this example, triggered a break in the opportunity cost of holding real money balances, and since the economy did not adjust immediately, a smooth-transition function was used to model the movement to a new money demand relationship.

Linking the forecast models used here, to those of Chapters 4 and 5, two points can be made. First, in theory and practice, some mechanistic correction mechanisms can produce accurate forecasts, without the need for

knowledge of the underlying DGP. however, over time the case for direct modeling of a break becomes stronger, as earlier forecast errors provide useful information about the nature of the break process.

Chapter 7

Forecast Density Evaluation with Structural Breaks

> I don't think we have failed, we have just found another way that doesn't work.
>
> ANDY ELLSON
> *British balloonist, after a failed attempt to fly around the world*

How can you judge when a forecast model has failed? In some cases, failure might be obvious – such as predicting fine weather, only to be hit by hurricane winds – but in other situations, it might be less clear cut, so more sophisticated methods of evaluation are necessary. However, the extent to which *they themselves* provide reliable indicators of forecast performance, is dependent on how robust they are to sources of forecast failure. This chapter considers one particular method of forecast evaluation that is relevant in the field of density forecasting, and examines how it performs in simulated cases of forecast failure. The aim of this exercise is to understand whether the evaluation method is robust to breaks in different elements of the density being forecast (such as its mean or variance). Then, the method is applied to the Bank of England's density forecasts of inflation, to see how the simulated cases compare to a real-world situation.

7.1 The Probability Integral Transform

In a forecasting exercise, the density of the data generating process $\mathsf{f}(\cdot)$ is likely to be unknown. However, this does not preclude the ability to evaluate how close a forecast is to the truth, thanks to the probability integral transform, or PIT, originally proposed by Rosenblatt (1952).

Although the $\mathsf{f}_{\mathsf{Y}_t}(y_t)$ is not observable, the PIT method exploits the fact that the *realisations* $\{y_t\}_{t=1}^n$ are. Thus the first step in the process of evaluation is to obtain the probability from the *forecast* density, of observing the *realized value* of Y_t, which is denoted z_t:

$$z_t = \int_{-\infty}^{y_t} \mathsf{g}_{\mathsf{Y},t-h}(u|\Omega_{t-h}, \theta_{t-h})\ du \equiv \mathsf{G}_{\mathsf{Y},t-h}(y_t|\cdot), \quad t = 1, \ldots, n \tag{7.1}$$

where $\mathsf{G}_{\mathsf{Y},t-h}(\cdot)$ is the forecast cumulative distribution function. Since z_t represents a probability, $z_t \in [0, 1]$. As Clements (2004) notes, (7.1) can be expressed in terms of the random variable Y_t rather than its realisations, such that

$$Z_t = \int_{-\infty}^{Y_t} \mathsf{g}_{\mathsf{Y},t-h}(u|\Omega_{t-h}, \theta_{t-h})\ du \equiv \mathsf{G}_{\mathsf{Y},t-h}(Y_t|\cdot).$$

Using a change of variable formula (see, e.g. Casella and Berger (2002, p.54)), the density of the PIT, denoted $\mathsf{q}_{\mathsf{Z},t}(z_t)$, is given by:

$$\begin{aligned}\mathsf{q}_{\mathsf{Z},t}(z_t) &= \mathsf{f}_{\mathsf{Y},t}\left(\mathsf{G}^{-1}_{\mathsf{Y},t-h}(z_t)\right)\left|\frac{\partial \mathsf{G}^{-1}_{\mathsf{Y},t-h}(Z_t)}{\partial Z_t}\right| \\ &= \frac{\mathsf{f}_{\mathsf{Y},t}\left(\mathsf{G}^{-1}_{\mathsf{Y},t-h}(z_t)\right)}{\mathsf{g}_{\mathsf{Y},t-h}\left(\mathsf{G}^{-1}_{\mathsf{Y},t-h}(z_t)\right)}\end{aligned}$$

If the forecast density is congruent, then $\mathsf{f}_{\mathsf{Y}_t}(\mathsf{G}^{-1}_{\mathsf{Y},t-h}(z_t)) = \mathsf{g}_{\mathsf{Y},t-h}(\mathsf{G}^{-1}_{\mathsf{Y},t-h}(z_t))$, and so $\mathsf{q}_{\mathsf{Z},t}(z_t) = 1$. Since it has already been established that $z_t \in [0, 1]$, this result is equivalent to writing $Z_t \sim \mathsf{U}[0, 1]$. In the case of one-step forecasts, where $h = 1$, there will be no overlap between each forecast horizon and the next; thus if all the forecasts are congruent, the sequence of forecasts should also be independently distributed, so $Z_t \sim \mathsf{UID}[0, 1]$. Longer forecast horizons can be evaluated, but for $h \geq 2$, the IID property is preserved only if the interval between each forecast is at least as long as the length of the horizon; otherwise there could be dependence in the PIT series by construction; this is discussed further in Clements (2005, Chapter 1).

This result is powerful, since evaluation of a forecast density is possible even if the DGP is unknown, as only the realisations $\{y_t\}_{t=1}^n$ are neces-

sary. Once a density forecast has been made, and the PITs constructed, the method of testing for forecast congruence with reality centres upon testing the sequence of the PITs, $\{z_t\}_{t=1}^n$, for independence and uniformity. Further, the approach is extendable to multivariate forecasts, as Diebold, Hahn and Tay (1999) discuss.

A number of methods exist to test for these properties: an overview in the context of density evaluation is provided by Corradi and Swanson (2006). Two broad approaches are possible; first, a suite of tests that evaluate the $\{z_t\}_{t=1}^n$ series directly, testing for uniformity and independence; and secondly, tests based on the transformation:

$$z_t^* = \Phi^{-1}(z_t) = \Phi^{-1}\left(\int_{-\infty}^{y_t} \mathsf{g}_{\mathsf{Y},t-h}(u|\Omega_{t-h}, \theta_{t-h})\ du\right),$$

where $\Phi^{-1}(\cdot)$ is the inverse of the standard normal distribution function. Addressing the first method, the direct uniformity tests typically assess the extent to which the empirical distribution function of the sequence $\{z_t\}_{t=1}^n$ is significantly different from the theoretical distribution function of a uniform random variable, which is a 45° line. Unfortunately, as Berkowitz (2001) observes, many of these non-parametric tests are notoriously data-intensive, suggesting the need for at least 1,000 observations for methods such as the Kolmogorov-Smirnoff (KS) test, or the Cramer-von Mises distance. In the context of inflation forecast density evaluation, for instance, where sample sizes of 30 are large, this is an infeasible requirement. Further, Noceti, Smith and Hodges (2003) suggests that the power of the KS test is lower than other alternatives, when mis-specification is present. In the context of testing for the independence of the PIT series, Mitchell (2005) and Hall and Mitchell (2007) also report poor performance of tests if dependence in the series is introduced.

With respect to the second approach, suggested by Berkowitz (2001), the main advantages of transforming the series lie in the fact that tests for normality are frequently easier to implement in practice, have been studied in detail before, and importantly, work well in small samples. Further,

as Berkowitz (2001, Proposition 2) demonstrates, this transformation, like the Rosenblatt transformation above, does not impose any distributional assumptions on the underlying data. Thus "inaccuracies in the density forecast will be preserved in the transformed data."[1]

Since many aspects of evaluation tests have been covered in detail in the literature (see, for example, Hall and Mitchell (2007)), the focus in this chapter is on the effects of structural breaks on a number of normality tests applied to the $\{z_t^*\}_{t=1}^n$ series.

In this regard, a number of methods are possible. A non-parametric approach is offered by a Doornik-Hansen test of the third and fourth moments, which has the power to detect departures from normality without imposing any distributional assumptions under the null hypothesis (see Doornik and Hansen (1994)). This has an advantage over alternative tests (see, for instance, Pearson, d'Agostino and Bowman (1977)), since it has good small-sample properties, which are useful for the simulations and applications performed here. The main limitation of this test, though, is the fact that departures from *standard* normality in the first two moments are not detected; since $\{z_t^*\}_{t=1}^n$ should have a zero mean and unit variance, this suggests the need for alternatives in addition.

Parametric tests that impose normality under the null have been suggested by Berkowitz (2001), either as a $\chi^2(1)$ test for independence, or a $\chi^2(3)$ test for zero mean, unit variance and independence. In this spirit, a simple likelihood ratio test for zero mean and unit variance is possible, which has a $\chi^2(2)$ distribution. These tests do have the power to detect non-normality through the first and second moments, but if a test is rejected, there is no indication of *why*; this is addressed by Bontemps and Meddahi (2005), who introduce a GMM approach to normality testing through moment conditions implied by a normal distribution. This has an advantage since the failure of a particular moment condition provides an indication of why a series might not be normal. Corradi and Swanson (2006) discuss several further tests and related issues.

[1] Berkowitz (2001), p. 467.

Table 7.1: SUMMARY OF TEST STATISTICS

Test	Distribution
Doornik-Hansen (Normality)	$\chi^2(2)$
$\mathsf{N}(0,1)$ (Likelihood ratio)	$\chi^2(2)$
$\mathsf{IN}(0,1)$ (Likelihood ratio)	$\chi^2(3)$
Chow breakpoint	$\mathsf{F}(41-t, t-2)$

NOTES: For the Chow breakpoint test, $t = M, \ldots, 40$, where M is the point in time at which the test is evaluated. In the case of recursive tests shown below, M starts at 15, rising to 39.

For the purposes of evaluation in this chapter, a diverse range of tests is used, and they are summarised in Table 7.1; a Doornik-Hansen (DH) test assesses general normality, whilst likelihood ratio tests examine normality along with the mean, variance and autoregressive parameters; a break-point Chow test evaluates the presence of structural breaks; and a test for serial correlation is used to detect dependence in the $\{z_t^*\}_{t=1}^n$ series. This evaluation strategy seems appropriate, since all the tests are computationally simple in simulation exercises, and can be applied to real-world data with similar ease.

7.2 Simulating the PIT

The discussion of the simulations performed here is divided into three parts: Section 7.2.1 first explains the components of the exercises, Section 7.2.2 presents and discusses results for simulations where there are no breaks, and Section 7.2.3 extends analysis to cases with breaks.

7.2.1 Simulation Exercises

The simulation exercises comprise two components: first, the specification of the DGP of Y_t and secondly, the forecast model $\mathsf{g}_{\mathsf{Y},\cdot}(\cdot)$; these are discussed in detail below, but it is useful first to outline the details of the simula-

tion process. After obtaining the sequence of transformed PITs $\{z_t^*\}$ from the exercises, for a sample size of $T = 40$, the evaluation tests described above are applied. This produces a test statistic for each, which can be compared to the critical value obtained from its relevant distribution. Thus for example, the statistic arising from the Doornik-Hansen normality test can be compared to the relevant critical value for a $p\%$ significance level from its $\chi^2(2)$ distribution. The benefit of a Monte Carlo simulation is that this process is repeated a large number of times (in these exercises, there are $100,000$ replications), and then the percentage of these replications for which the test statistic is significant, is reported.

7.2.1.1 Specifying the Data Generating Process

Using Ox, the data generating process is simulated, so that a sequence of observations is generated, and then used to derive the PIT sequence, using the forecast density specified in Section 7.2.1.2. In the exercises without breaks, four DGPs are simulated:

(a). $y_t \sim \mathsf{NID}(0, 1)$.

(b). $y_t \sim \mathsf{NID}\left(\frac{1}{2}, 1\right)$

(c). $y_t \sim \mathsf{NID}\left(0, \frac{1}{2}\right)$

(d). $y_t \sim \mathsf{N}(0, 1)$, but $y_t | y_{t-1} \sim \mathsf{N}\left(0.5y_{t-1}, \frac{3}{4}\right)$, where $y_0 = 0$

Case (a) is the baseline case against which the other exercises can be compared; (b) alters the mean, and (c) the variance of the DGP, without affecting the independence property, whilst (d) introduces a first order autoregressive structure into the Y_t process.

When breaks are introduced, the DGP changes at $t = 25$, which is a suitable point close to the mid-point of the sample. Two break cases are simulated: first, a shift in the mean of the DGP, and secondly, a shift in both the mean and variance, to reproduce a common observation in macroeconomic time series, that both the mean and variance of a series such as inflation shift together. In the presence of breaks, forecasting models can

be affected in different ways, depending on their properties. For example, Chapter 4 considered 'robust' models that may break down at the point of a break, but correct themselves rapidly in its aftermath. Thus it is useful to consider the simulated breaks in two scenarios: in the first, the break at $t = 25$ is permanent, so that it persists for the remainder of the sample. The second scenario, in contrast, sees the DGP shift *back* to its original state two periods after the break, simulating a situation in which a robust model would fail immediately following the break, but then set itself back on track. In total, four cases are considered:

(I) i.

$$y_t \sim \begin{cases} \mathsf{NID}\,(0,1) & t < 25 \\ \mathsf{NID}\left(\frac{1}{2},1\right) & t \geq 25 \end{cases}$$

ii.

$$y_t \sim \begin{cases} \mathsf{NID}\,(0,1) & t < 25 \\ \mathsf{NID}\left(\frac{1}{2},1\frac{1}{2}\right) & t \geq 25 \end{cases}$$

(II) i.

$$y_t \sim \begin{cases} \mathsf{NID}\,(0,1) & t < 25 \\ \mathsf{NID}\left(\frac{1}{2},1\right) & 25 \leq t \leq 26 \\ \mathsf{NID}\,(0,1) & t \geq 27 \end{cases}$$

ii.

$$y_t \sim \begin{cases} \mathsf{NID}\,(0,1) & t < 25 \\ \mathsf{NID}\left(\frac{1}{2},1\frac{1}{2}\right) & 25 \leq t \leq 26 \\ \mathsf{NID}\,(0,1) & t \geq 27 \end{cases}$$

7.2.1.2 The forecasting model and the PIT

For all specifications of the DGP listed above, the forecast model is given by:

$$\mathsf{g}_{\mathsf{Y},T-h}(y_T|\Omega_{T-h},\theta_{T-h}) \overset{\mathsf{D}}{=} \mathsf{NID}\,(0,1)\,.$$

Using this forecast density, the PIT sequence can be constructed according to (7.1) and from this, the normal transformation is used to derive $\{z_t^*\}$, which can then be tested accordingly.

7.2.2 Simulations without breaks

In view of the sensitivity of statistical tests to the small sample size of the simulations, it is helpful to start by looking at the empirical distribution of the probability integral transforms, shown in Figure 7.1. The left hand column shows the frequency plots of the PITs $\{z_t\}$, and on the right, the corresponding quantile plot; each row corresponds to cases (a) to (d) above. Under a congruent density forecast, the former should be uniformly distributed over the unit interval [0,1], and the latter should be close to the 45° line which corresponds to the quantiles of a uniform random variable.

Four observations can be made about Figure 7.1. The first row shows the distribution when the model is well-specified, and the PIT sequence z_t clearly has a uniform frequency distribution and quantile plot.

Secondly, the distribution of the PIT in the second row is consistent with the mean of the DGP being higher than that of the forecast density. If $\{z_t\}_{t=1}^n$ is generated by a Gaussian process with mean $\frac{1}{2}$ and unit variance, then only approximately 30% of the realisations fall below the mean of the forecast distribution (which is zero in this case), rather than the 50% expected with a congruent forecast.

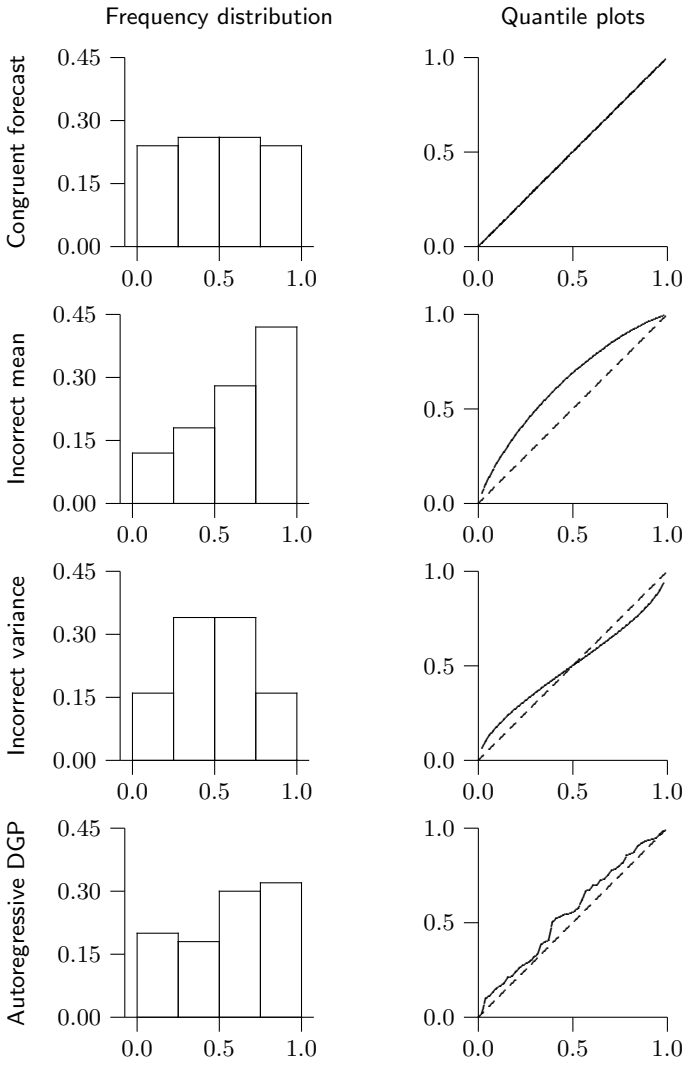
The third row demonstrates the effect of the forecast model overstating the true variance, as the PIT values are concentrated around the mean of $\frac{1}{2}$, and not sufficiently spread out.

Finally, the fourth row suggests that graphical inspection of the PIT empirical density is not particularly informative in small samples in the case of incongruent densities, when the DGP and forecast *share the same mean*. Even though this case is mis-specified, it is difficult to determine this from Figure 7.1, for the quantile plot of one replication chosen at random.[2]

Figure 7.2 shows rejection frequencies of the null hypothesis of an $\mathsf{N}(0,1)$

[2]The quantile plot in the fourth row of the figure reflects one replication, rather than the average over all 100,000 replications, because the intention here is to illustrate the effect of the forecast density not matching the conditional DGP density. Taking the average over many replications produces a PIT distribution that reflects the *unconditional* distribution of the DGP, which is the same as that of the forecast, rather than the *conditional* distribution, which is of interest here.

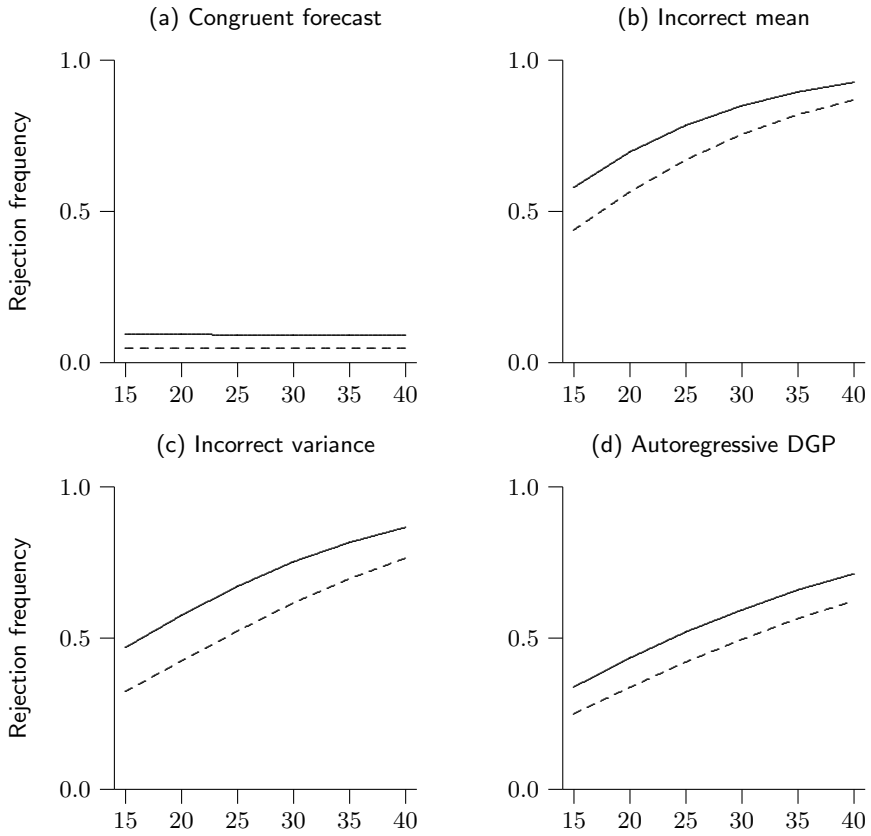
Figure 7.1: PROBABILITY INTEGRAL TRANSFORM DISTRIBUTION



NOTES: Panel rows show simulation results corresponding to the four cases (a) to (d) explained in the text: (a), when the forecast density is congruent; (b), when the forecast density has the wrong mean; (c), when the forecast density has the wrong variance, and; (d), when the DGP has an autoregressive structure that is not captured by the forecast model. The left hand panel is the frequency plot of the $\{z_t\}$ series, whilst the right hand panel shows the quantile plot. Note that the first three quantile plots show the average quantiles over $100,000$ replications, whilst the fourth row shows the quantile plot of an individual replication chosen at random. *Ox code reference:* `PIT-simulations.ox`

distribution in the transformed PIT sequences $\{z_t^*\}_{t=1}^k$, for $k = 15, \ldots, 40$; thus it provides a profile of the test behaviour as the sample size increases.

Figure 7.2: DENSITY EVALUATION TESTS



NOTES: Panels show the proportion of cases over 100,000 replications in which the null hypothesis of a particular test was rejected. The tests were applied to the $\{z_t^*\}$ series, relating to four cases: (a), (b) and (c) are tests for a $\mathsf{N}(0,1)$ distribution under the null hypothesis, whilst (d) is a test for an $\mathsf{IN}(0,1)$ distribution. Recursive estimation performed for sample size $T = 15$ to 40, in steps of 5. *Ox code reference:* `PITsim.ox`

In all four panels, the test is well-behaved, since it correctly detects the presence of different forms of mis-specification, although it cannot distinguish between the underlying causes.

Case (d), in the bottom right panel is of particular interest. Introducing an autoregressive structure in the DGP of Y_t has the effect of violating the independence of the PIT. This is evident in the panel, since the test evaluates whether the PIT has a normal distribution with zero mean, unit variance, and an autoregressive coefficient of zero. The precise cause of this violation is that, even though the forecast density is the same as the *uncon-*

ditional distribution of Y_t, the PIT method implemented here evaluates the *conditional* distribution; thus the correct conditional density in this case is given by

$$\mathsf{g}_{\mathsf{Y},t-h}(y_t|\Omega_{t-h},\theta) \overset{\mathsf{D}}{=} \mathsf{N}(0.5y_{t-1}, 0.75),$$

rather than $\mathsf{N}(0,1)$. This raises an important implication for the evaluation of any models in which serial dependence might be present in the DGP: evaluation in this case needs to be performed using the conditional distribution.

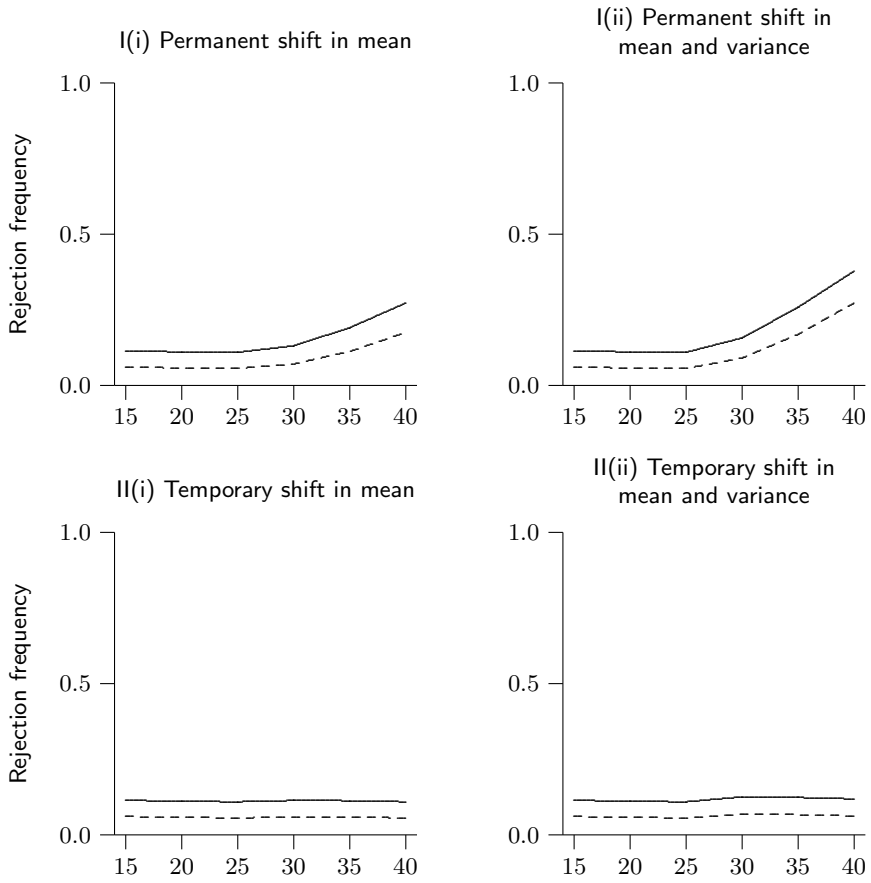
7.2.3 Simulations with breaks

Simulating the evaluation tests in the presence of breaks raises a number of important issues. Figure 7.3 shows the rejection frequencies of normality tests for the case of a permanent break (I) in the upper row, and a temporary shift (II) in the bottom row. The left- and right-hand columns represent breaks in the mean and variance respectively. As the figure demonstrates, both breaks are detected when they are permanent, although the rejection frequency remains low for periods closely following the break. Interestingly, whether the break is permanent or transitory, neither shift is conclusively detected by Chow break-point tests (results not reported). This is not true for larger breaks, though, and Figure 7.4 demonstrates the effect of a shift in the DGP mean of two standard deviations, for both a permanent and temporary break.

In this case, whether the break is permanent or temporary has a significant effect on the results of the recursive tests. For a shift in the density lasting only two periods, normality in the series is rejected for *all* subsequent sample periods (i.e. $1,\ldots,j$ for $j > 25$). A permanent shift, though, is consistently detected by the Chow test. This is important, since it suggests that the normality test alone is insufficient in evaluating the PIT sequence when a break occurs: a temporary break need not suggest permanent forecast failure, and so the fact that the normality test *does* suggest this, could be problematic.

These results raise a number of interesting implications. First, they

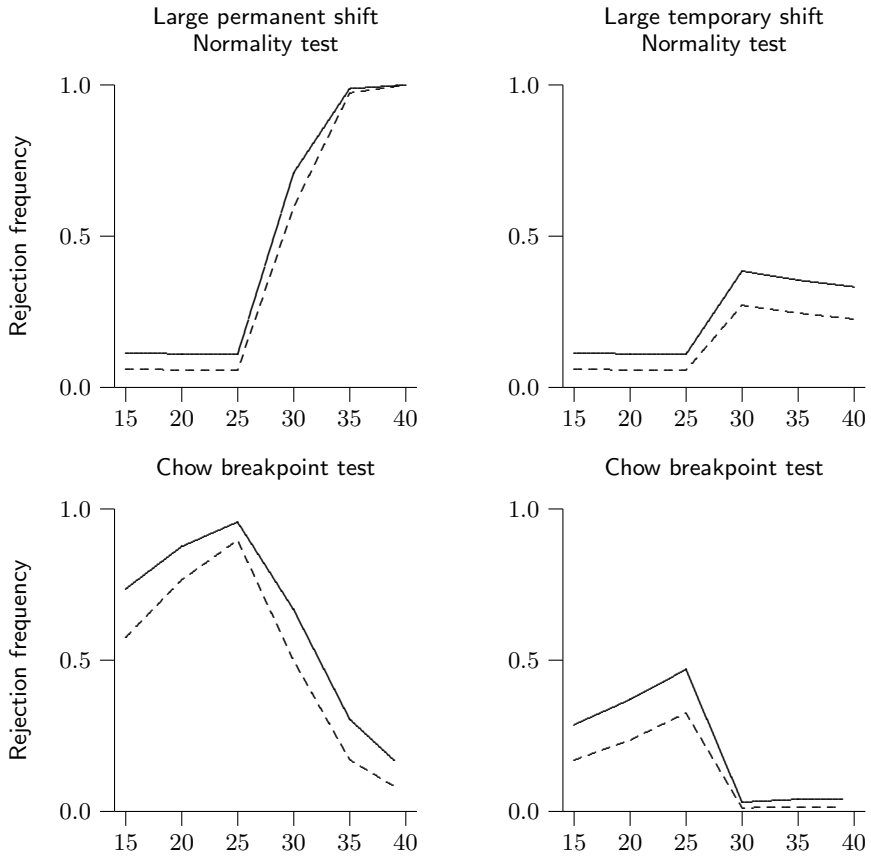
Figure 7.3: PIT TESTS WITH SIMULATED BREAKS



NOTES: Panels show the proportion of the 100,000 simulations for each sample size in which the null hypothesis was rejected. In all cases, the null hypothesis was of a PIT sequence having a $\mathsf{N}(0,1)$ distribution. In the left-hand *column*, the true model has a shift in the mean of $\frac{1}{2}$ a standard deviation at time $t = 25$, and in the right-hand *column*, both the mean and variance increase by $\frac{1}{2}$ of one standard deviation. In the upper *row*, the break in each case lasts for all time periods after $t = 25$, whilst in the lower *row*, the break lasts for only two periods, after which the model reverts to its pre-break state. *Ox code reference:* `PITsim.ox`

reaffirm the need for a *range* of tests to evaluate the sequence of PITs: there does not seem to be one specific test that can consistently detect unanticipated breaks in the forecast density. As Clements (2004) discusses, Thompson (2004) derives a portmanteau test for a joint test of uniformity, first-moment independence and unpredictable variance, and so a profitable direction for future research would be to analyse the effect of breaks on this aggregate measure.

Figure 7.4: PIT TESTS WITH LARGE SIMULATED BREAKS



NOTES: Panels show the proportion of the 100,000 simulations for each sample size in which the null hypothesis was rejected. In all cases, the null hypothesis was of a PIT sequence having a $\mathsf{N}(0,1)$ distribution. In all simulations, there is a break in the mean of the data generating process of 2 standard deviations at time $t = 25$; the left-hand *column* report simulations in which this break is permanent, whilst the right-hand *column* reports results when the break lasts for two periods, after which the model reverts to its pre-break state. The upper *row* shows tests for an $\mathsf{N}(0,1)$ distribution, and the lower *row* reports Chow breakpoint tests. *Ox code reference:* `PITsim.ox`, `PIT-PCNAIVE-Iilarge.ox`, `PIT-PCNAIVE-IIilarge.ox`

Secondly, the failure to detect small breaks raises problems with the reliability of evaluation tests in small-sample forecasting exercises. Even though a small permanent break in the mean is detected, the rejection frequency for a 10% significance level only increases to 25% for the full sample (of 40 observations). In context, a break of half a standard deviation in the Bank of England's inflation forecast would could have a magnitude of up to 0.4

percentage points of inflation; since the Bank's remit from the Government is to target inflation of 2%, within a symmetric band of ±1%, such a shift would not be insignificant.

A further conceptual point relates to the behaviour of the PIT series when there is a break for only two periods – that is, a 'blip' in the mean of the true density. In principle, it might be desirable for the evaluation method *itself* to be robust to such shifts, since they could well be unavoidable, and not signify any kind of forecast failure in a particular model. In this respect, the evidence suggests that the PIT as a method of evaluating density forecasts is robust if the shifts are small (the top left panel in Figure 7.3), but less so if larger shifts occur (the top right panel in Figure 7.4). In a sense, requiring an evaluation method to detect discrepancies when it 'needs to' (that is, when a break has actually taken place), but at the same time, expect it to be robust when breaks are temporary, may be unreasonable. However, this argument ignores an important need to know how these evaluation tests behave in various structural break scenarios; further, there remains a need for a forecast evaluation method that is robust to one-off breaks, since these may be unavoidable even for the best performing forecast models. As the next section demonstrates, many of the issues raised above are relevant when applying the PIT method to a sequence of published forecasts.

7.3 Forecast Evaluation and UK Inflation

The Bank of England is one of several Central Banks to publish a detailed breakdown of its inflation forecast. Using the evaluation techniques established above, it is therefore possible to assess to some extent, whether the Bank's published forecasts match the true density of inflation. The motivation for this lies once again within a decision-theoretic environment: regardless of the loss function of the Bank's Monetary Policy Committee (MPC), a congruent forecast is weakly superior to all other incorrect forecasts. Several studies have evaluated the Bank's inflation forecasts, including Clements (2004), Wallis (1999, 2004) and Hall and Mitchell (2007); however, no study has examined the series of PITs derived from the forecasts to assess the

presence of breaks.

7.3.1 The Bank of England's inflation forecast

The model used by the Bank of England to represent the MPC's uncertainty about inflation over its two-year forecast horizon is discussed in detail in Britton *et al.* (1998) and the references cited above; thus this section contains only a summary of its relevant features.

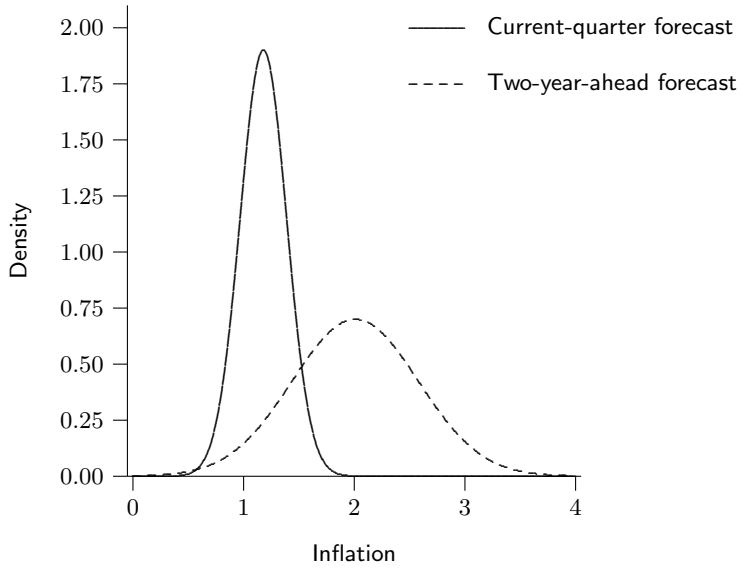
Since 1996, the Bank's inflation forecast has been published explicitly in the form of a probability distribution, as Britton *et al.* (1998) observe. Graphically, this is presented as a series of probability intervals around a central scenario which corresponds to the modal outcome in the distribution. Analytically, the probability distribution is given by a two-piece, or split, normal distribution. The rationale for this formulation is that it provides a straightforward (analytical) method of characterising the fact that 'risks' to the central inflation projection (i.e. the mode) may be *weighted* on the upside or downside – equivalently, that the probability distribution is not symmetric around the mode, and therefore possesses skew.

Formally this is expressed by dividing the distribution into two halves around a common mode μ^d. Thus for realisations of inflation below μ^d, the forecast is normally distributed with mean μ^d and variance σ_1^2; conversely for realisations above μ^d, inflation has a variance σ_2^2. Thus with x denoting inflation, the probability distribution is given by (see John (1982)):

$$\mathsf{D}_{\mathsf{x}}(x) = \begin{cases} A \exp\left\{\frac{-1}{2\sigma_1^2}(x-\mu^d)^2\right\} & x \leq \mu^d \\ A \exp\left\{\frac{-1}{2\sigma_2^2}(x-\mu^d)^2\right\} & x \geq \mu^d \end{cases} \tag{7.2}$$

where $A = 2\left(\sqrt{2\pi}(\sigma_1+\sigma_2)\right)^{-1}$ is a normalisation constant that ensures the combined two-piece distribution has a common value $\mathsf{D}_{\mathsf{x}}(x) = A$ at μ^d. The relation of the mean to the mode is given by the direction of the distribution's skew: thus if $\sigma_1 > \sigma_2$, then more than half of the probability mass will lie below the mode, resulting in a negatively skewed distribution, as shown in Figure 7.5. Consequently, mean<median<mode; if $\sigma_2 > \sigma_1$, the

Figure 7.5: BANK OF ENGLAND INFLATION DENSITY FORECASTS



NOTES: Densities are taken from the Bank of England's forecast densities published with the *Inflation Report*. Both forecasts relate to inflation in 2004Q3, based on interest rates following the path implied by market expectations.

distribution is positively skewed and so all the features above are reversed accordingly.

As Clements (2004) notes, the cumulative distribution function for the two-piece model is given by

$$\mathsf{F}_{\mathsf{X}}(x) = \mathsf{P}(X \leq x) = \begin{cases} \frac{2\sigma_1}{\sigma_1+\sigma_2}\Phi\left(\frac{x-\mu^d}{\sigma_1}\right) & x \leq \mu^d \\ \frac{2\sigma_2}{\sigma_1+\sigma_2}\Phi\left(\frac{x-\mu^d}{\sigma_2}\right) + \frac{\sigma_1-\sigma_2}{\sigma_1+\sigma_2} & x \geq \mu^d \end{cases} \tag{7.3}$$

where as before, $\Phi(\cdot)$ denotes the cumulative distribution function of a standard normal distribution; from this, it is clear that in the symmetric case $\sigma = \sigma_1 = \sigma_2$, the distribution collapses to a conventional normal distribution $\mathsf{N}(\mu^d, \sigma^2)$.

A final point to note, following Clements, is that each *Inflation Report* publishes two sets of projections: one in which interest rates are assumed to be constant over the forecast horizon (referred to here as the constant interest rate forecast), and one in which interest rates are assumed to follow

market expectations (the market interest rate forecast). The rationale for this lies in the fact that the MPC's beliefs about the future rate of interest are in themselves sensitive information, and subject to change before its monthly meeting. Thus by publishing a forecast of inflation based on its own assumptions regarding the path of interest rates, the Bank would risk 'revealing its hand' early, potentially undermining the impact of changes at a later date. In deriving an interest rate forecast based on a measure of the *market's* expectations, this particular problem is circumvented, as the information set needed to derive the series is not restricted to the MPC or the Bank. However, it does raise a separate problem of how to derive market expectations: this is discussed in Pagan (2002) and MPC (2004).

The *a priori* expectation with regard to the congruence of each projection is unclear, since the performance of the market expectation-based scenario depends upon the method of deriving the interest rate forecast. However, in general it might be true that the fixed interest rate assumption is unrealistic, and so a flexible interest rate assumption would produce a 'better' projection, especially for longer-horizon forecasts.

7.3.2 Bank of England forecast parameters

As discussed in Wallis (2004), the inflation forecast parameters published by the Bank of England[3] report the mode, median and mean of each forecast, in addition to a measure of uncertainty, denoted σ, and skew, denoted s and given by the difference between the mean and the mode of the distribution. For each *Inflation Report*, parameters are provided for the current, and subsequent eight quarters, corresponding to a sequence of one to nine-step ahead forecasts. In this exercise, following Clements, three forecast horizons are chosen for evaluation: one-step (i.e. current quarter), five-step (one year ahead) and nine-step (two years ahead). Using a recent Bank spreadsheet,[4] 42 current quarter forecasts are available, 38 one year ahead forecasts, and 34, two years ahead. Although the density forecasts made

[3] At http://www.bankofengland.co.uk/publications/inflationreport/irprobab.htm.
[4] Downloaded February 2008.

from 2004 onwards, relate to the new CPI measure of inflation, rather than the earlier RPIX measure, this should have no effect on evaluation, since the appropriate measure is used in the sequence of realisations of inflation.

Following Wallis, the published statistics are used to derive the parameters in the two-piece distribution. Defining s^* as a standardised skew, where $s^* = \frac{s}{\sigma}$, then a bounded skew, denoted γ, is given by:[5]

$$\gamma^2 = 1 - 4\left(\frac{\sqrt{1+\pi s^{*2}} - 1}{\pi s^{*2}}\right)^2,$$

where, as Wallis observes, 'γ takes the sign – positive, negative, or zero – of s.' Then, given γ and σ:

$$\sigma_1^2 = \frac{\sigma^2}{1+\gamma} \quad \sigma_2^2 = \frac{\sigma^2}{1-\gamma}.$$

Characterising the two-piece normal distribution with these derived parameter values, it is then possible to construct the realisations of the probability integral transform as explained in Section 7.1.

7.3.3 Forecast evaluation

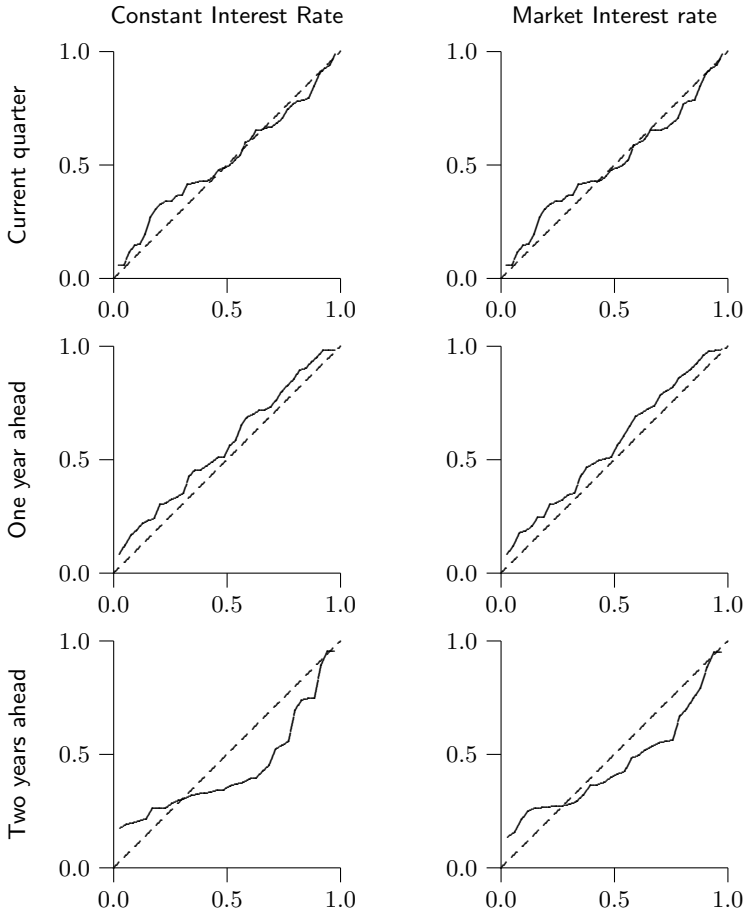
Following the format established in the simulation exercises, the evaluation of the Bank's density forecasts begins with a graph of the empirical distribution of the probability integral transform.[6] With a congruent forecast, each quantile plot should be close to the straight line, and the results for each forecast horizon are shown in Figure 7.6. Whilst the current quarter and one year ahead forecasts do appear to perform well, both the constant and market interest rate forecasts indicate significant departure from the 45° line at a two year horizon.

Further recursive statistical tests are shown in Figure 7.7, where the first row reports DH p-values for the market interest rate forecasts; the second row reports Chow breakpoint test results for the market interest rate forecasts, and the third row shows the p-values for a $\chi^2(2)$ likelihood

[5] See Wallis (2004, Box A.).

[6] All computations and tests were performed using *Ox* (see Doornik (2007)).

Figure 7.6: BANK OF ENGLAND INFLATION FORECAST EVALUATION: PROBABILITY INTEGRAL TRANSFORM QUANTILES

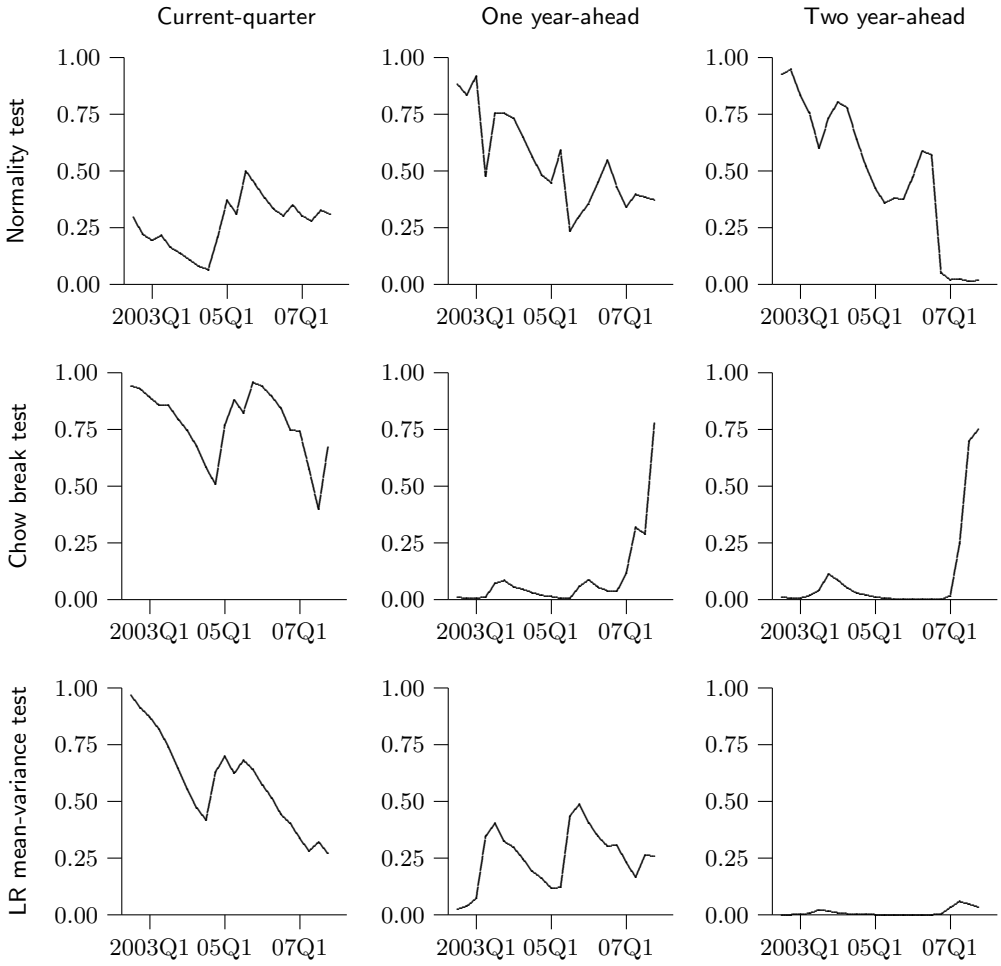


NOTES: Panels show quantile plots of the probability integral transform for each forecast. Rows relate to the stated forecast horizon, and the left- and right-hand columns relate respectively to inflation forecasts made under the assumption of constant interest rates, and market expectations of interest rates. *Ox code reference:* `BankTestPIT.ox`

ratio test of normality with zero mean and unit variance, performed on the market interest rate forecasts. Columns indicate the length of the forecast horizon, which is given at the top of each.

Several points can be made about the results. First, the one-step (i.e. current quarter) results seem to be accurate, with all tests failing to detect relevant mis-specification. Secondly, whilst the one year-ahead forecasts perform well on both DH and LR tests, there does seem to be evidence of

Figure 7.7: FORECAST EVALUATION STATISTICS



NOTES: The first row of panels refer to DH tests of normality for market-based interest rate inflation forecasts, at three horizons; the second row corresponds to Chow structural break tests for the market-based interest rate forecast, and the third refers to a likelihood ratio test for zero mean and unit variance. All rows report p-values, against null hypotheses of normality, structural stability, and zero mean and unit variance, respectively. The time axis reflects the forecast horizon. *Ox code reference:* `BankTestPIT.ox`

structural instability over the period, which only calmed down towards the end of the sample.

Further, the two year ahead forecasts suggest some degree of forecast inaccuracy. First, the DH test points towards non-normality in the PIT sequence as 2007 progressed, although earlier forecasts were more consistent

with the normal null hypothesis. Secondly, there is significant evidence of structural instability, and the likelihood ratio test points to decisive rejection of the $\mathsf{N}(0,1)$ hypothesis. It would seem as though there is evidence of a systematic deterioration in performance moving into 2007, which could reflect changes in the behaviour of inflation that the Bank of England did not forecast accurately in 2005, when the relevant two year-ahead forecasts were produced.

The implications of these results for the decision-making process in the MPC, though, are unclear. On one hand, the performance of the Bank's two year-ahead inflation forecasts is crucial, given the time lag between interest rate decisions and their overall effect on the economy. Thus an inaccurate prediction of inflation over this period could lead to incorrect policy decisions. However, the representation of forecast uncertainty through a two-piece normal distribution is only an *approximation* that provides an intuitive and understandable demonstration of the general risks surrounding an inflation projection, and any *asymmetry* of risks. It does not claim to represent a full description of the model, data and estimation uncertainties surrounding the inflation forecast. Further, the decision-making process itself is based on discussions of the MPC, which themselves shape the forecast of inflation and the policy decision simultaneously: there is no mechanistic link running from a forecast to a decision. Thus evaluating the quality of forecasts does not amount to evaluating the quality of decisions made.

7.4 Implications and Concluding Comments

In conclusion, the probability integral transform offers an appealing and straightforward method of evaluating density forecasts for their statistical value. Even without knowledge of the data generating process, it is still possible, in a wide range of cases, to assess how closely a forecast matches the true density of a random variable. Further, the approach has natural links to the intellectual framework of decision theory in econometrics, and as such, it has intuitive appeal that provides a clear motivation for examining density forecasts: a congruent forecast weakly dominates all other forecasts a

decision-maker might use; and so, where forecasts are used to make decisions, their accuracy ought to be assessed.

However, the context in which forecasts are evaluated in this way, whilst providing a clear motivation, also imposes limits on the conclusions and implications that can be drawn. Thus the presence of structural breaks, for example, poses a number of challenges. On a simple level, breaks pose a challenge to standard measures of evaluation, since small shifts are not always detected by conventional normality tests, for instance. This would suggest that a prudent evaluation strategy should include a suite of tests, rather than relying on only one test type.[7]

More fundamentally, there are serious limits to the informativeness of density evaluation in this way. Whilst the PIT as a method of evaluation can provide useful information about the accuracy of a particular sequence of forecasts in a retrospective, *ex post* sense, and might therefore be useful to a decision-maker wishing to evaluate the quality of inputs into the decision problem, as a guide for choosing a forecast model to use in the future, it has very little power. Structural breaks can easily render a well-fitting, congruent model essentially useless if it is not robust to such shifts. Vulnerable models have the main characteristic that a structural break entails them systematic bias, inducing permanent forecast failure. Thus even models that *are* an accurate representation of in-sample data, and do perform well in an *ex post* forecast evaluation exercise, may well fail if there is a break during the forecast horizon.

[7]This view is shared by Hall and Mitchell (2007), who follow an 'eclectic' approach to testing.

Chapter 8

Conclusions

> Reports that say that something hasn't happened are always interesting to me, because as we know, there are known knowns; there are things we know we know. We also know there are known unknowns; that is to say we know there are some things we do not know. But there are also unknown unknowns – the ones we don't know we don't know
>
> DONALD RUMSFELD, FORMER US SECRETARY OF DEFENSE

This thesis indicates that an econometrician cannot ignore structural variation in an economy. Variation across individuals or over time, presents a significant challenge to conventional techniques for modeling and forecasting. Cross-section heterogeneity (for example, across countries or industries) can generate biased and inconsistent results in regressions that take an aggregate perspective; and mean breaks over time can lead to systematic forecast failure in some models, and even evaluation methods suffer from shifts.

In a cross-section dimension, the underlying intuition is that factors like technology, tastes and production functions have a large influence on how an industry behaves in the face of a rise in demand, for instance, so it is reasonable to believe that structural differences across industries should lead them to respond differently in the face of the same shocks.

Chapter 2 examined the import and export behaviour of a set of manufacturing industries in several developed countries. Disaggregating the data from an OECD- or country-wide level revealed that there is significant variation in the dependent and explanatory variables, that is disguised by aggregation. Further, empirical analysis found that an assumption of common

elasticities for all groups is not supported by the data; rather, a picture emerges of variation across industries in long-run trade behaviour.

The policy implications of these results are twofold. First, differences across industries in trading patterns with partner countries mean that a given change in an aggregate exchange rate can have different effects on trade. Thus forecasts of imports and exports should take this into account. Secondly, where policies are targeted at specific sectors, or types of industry (such as research and development spending aimed at hi-tech industries), accurate estimates of elasticities are important, as factors like innovation and trade performance are interdependent.

The challenge to forecast models posed by mean shifts highlights two themes that have underpinned the second Part of this thesis. In the first place, breaks in the mean of a process are a major source of forecast failure as they introduce bias in forecast models. Unless a type of monitoring device is used to predict when an equilibrium shift occurs (and therefore allow for pre-emptive correction), this bias is inevitable when a break takes place. The second theme follows directly from this, and revolves around the fact that the breaks studied here were unpredictable with respect to any available information set. This effectively precludes the use of any monitoring device, and explains why forecast failure at the point of a break is unavoidable.

With this in mind, the search for a robust model – one which does not suffer systematic failure after a break – must concentrate on the best response to a shift in its immediate aftermath. This motivated the robust models discussed in Chapters 4, 5 and 6, which sought to incorporate the information revealed at the point of a break in such a way that future forecasts would not suffer from systematic bias. In assessing the success of these methods, it is helpful to distinguish between 'mechanistic' (or automatic) correction mechanisms such as the double-differenced device, the differences VEqCM and the intercept-correction model, and more direct attempts to model the break process such as the ogive weight function in Chapter 6.

When breaks are likely to happen, a mechanistic device provides insurance against their effects in the periods after they take place, at a cost of

raising the forecast uncertainty whether or not there are shifts. In some situations, where models are mildly mis-specified (as defined in Chapter 4), this cost is attenuated, since the robust mechanism can also reduce the impact of mis-specification. However, in other cases, most notably when there is measurement error in the variable being forecast, robust devices can make matters worse, by attempting to correct for breaks when in reality none has taken place.

In contrast to the automatic devices that operate in a pre-determined way, an active approach to forecast adjustment after a break is also possible, and demonstrated in Chapter 6. Intuitively, a method in which a break process is modeled explicitly ought to perform well, as new information that emerges in forecast errors is used immediately to estimate the parameters in the adjustment process. However, as the chapter showed, even if the break point and the form of this learning function are known, the fact that parameters must be estimated means that for the first two periods after a break, mechanistic devices dominate the active response. As the empirical example suggested, only after three periods does the learning-adjusted model produce better forecasts, although this demonstrates that some form of correction pays dividends, as the mis-specified, uncorrected, model continues to suffer from forecast failure.

Turning to the evaluation method examined in Chapter 7, the message emerging from simulation exercises was that the tests had difficulty differentiating between temporary and permanent breaks, in the face of a large shift. As the preceding discussion has noted that some kind of forecast failure in the aftermath of an unpredictable break is inevitable, this result with evaluation methods provides some cause for concern.

Avenues for future research

With regard to Chapter 2, there are a number of directions to explore. The first is to extend the sample to cover a longer time period, so that structural variation in both cross-section and time-series dimensions can be explored, via techniques for analysing structural breaks in panel datasets.

The question that arises immediately, though, is how the time coverage of the data can be extended. One option is to wait for the forthcoming release of an updated STAN database, whilst another is to limit the country coverage to those for which a longer time span of data is available (for instance, starting in 1970). The former choice is constrained by uncertainty over future data releases, whilst the latter involves a compromise between the country and time scope of any empirical work.

A second route to explore is the role of other influences in determining import and export production capacity. The discussion in Chapter 2 highlighted that some factors, such as the level of technological progress, affected export performance, and their roles could be explored more thoroughly by including additional terms, such as a measure of innovation, in the long-run trade relationships. Driver and Wren-Lewis (2005) include cumulative investment to capture the effect of innovation on export market share in a panel of countries and further work could be done in our industry-country panel using the OECD Analytical Business Enterprise Research and Development (ANBERD) database, which matches the coverage of STAN (see OECD (2006*a*)). The role of pricing behaviour by exporters and domestic producers can also be incorporated into the framework developed here. Import-price pass-through, for example, was not addressed in Chapter 2, and yet studies have found that import prices can have an effect on domestic prices (see, for example, Nielsen and Bowdler (2006)). Thus an avenue for further work could look at the effect of this on elasticities.

With respect to the time-series forecasting models used in Part 2, further developments in both empirical and theoretical directions are possible. The framework developed by Castle *et al.* (2010) can be applied to wider empirical settings, in particular those where hysteresis mechanisms might help to explain why seemingly short-term shocks had long-term consequences. Examples in this realm include instances of trade hysteresis (following exchange rate shocks), linking in with the panel approach set out above; the pattern of UK and US unemployment in the interwar period; and trends in the housing market, where changes in expectations can cause bubbles of

overvaluation which then collapse.

Models of impulse saturation developed by Hendry, Johansen and Santos (2008) can be used in methods of forecast evaluation, to help distinguish between temporary and permanent breaks in the data generating process, which may provide the PIT-based evaluation procedure with some robustness in the face of shifts. Finally, the work on intercept correction models when there is measurement error could be extended into a state-space framework, to distinguish more explicitly between errors arising from structural change, and those emerging from data mis-measurement.

In conclusion, this thesis has found that structural variation across individuals and over time should not be ignored, and it explored a number of approaches that can deliver more reliable results, even if there is cross-sectional heterogeneity, or there are structural breaks. However, these solutions have limitations, and there are a number of directions for future research to improve them. In practice, structural differences will continue to occur, and so the task ahead is to study how, and why, it happens.

Bibliography

Balassa, B. (1979), 'Export Composition and Export Performance in the Industrial Countries, 1953-71', *Review of Economics and Statistics*, **61** (4), 604–607.

Baldwin, R. E. (1988), 'Hysteresis in Import Prices: The Beachhead Effect', *American Economic Review*, **78** (4), 773–785.

Baldwin, R. E. and Lyons, R. K. (1994), 'Exchange rate hysteresis? Large versus small policy misalignments', *European Economic Review*, **38** (1), 1–22.

Banerjee, A., Dolado, J., Galbraith, J. W. and Hendry, D. F. (1993), *Co-integration, Error Correction and the Econometric Analysis of Non-Stationary Data*, Oxford, Oxford University Press.

Banerjee, A., Marcellino, M. and Osbat, C. (2004), 'Some cautions on the use of panel methods for integrated series of macroeconomic data', *Econometrics Journal*, **7**, 322–340.

Berkowitz, J. (2001), 'Testing density forecasts, with applications to risk management', *Journal of Business & Economic Statistics*, **19** (4), 465–474.

Bontemps, C. and Meddahi, N. (2005), 'Testing normality: a GMM approach', *Journal of Econometrics*, **124** (1), 149–186.

Boswijk, H. P. (1992), *Cointegration, Identification and Exogeneity*, Vol. 37 of *Tinbergen Institute Research Series*, Amsterdam, Thesis Publishers.

Boswijk, H. P. and Doornik, J. A. (2004), 'Identifying, estimating and testing restricted cointegrated systems: An overview', *Statistica Neerlandica*, **58** (4), 440–465.

Breitung, J. (2005), 'A parametric approach to the estimation of cointegration vectors in panel data', *Econometric Reviews*, **24** (2), 151–173.

Breitung, J. and Das, S. (2005), 'Panel unit root tests under cross sectional dependence', *Statistica Neerlandica*, **59** (4), 414–433.

Breitung, J. and Pesaran, M. H. (2008), 'Unit Roots and Cointegration in Panels', in L. Matyas and P. Sevestre, eds, *The Econometrics of Panel Data, 3rd ed.*, Kluwer Academic Press, Amsterdam.

Britton, E., Fisher, P. and Whitley, J. (1998), 'The *Inflation Report* projections: understanding the fan chart', *Bank of England Quarterly Bulletin*, **38** (1), 30–37.

Cameron, G., Proudman, J. and Redding, S. (2005), 'Technological Convergence, R&D, Trade and Productivity Growth', *European Economic Review*, **49** (3), 775–807.

Carlin, W., Glyn, A. and van Reenen, J. (2001), 'Export market performance in OECD countris: An examination of the role of cost competitiveness', *Economic Journal*, **111** (468), 128–162.

Casella, G. and Berger, R. L. (2002), *Statistical Inference*, 2nd edn, Duxbury.

Castle, J. L., Fawcett, N. W. P. and Hendry, D. F. (2010), 'Forecasting, Structural Breaks and Non-linearities', *Journal of Econometrics*, DOI:10.1016/j.jeconom.2010.03.004.

Chamberlain, G. (2000), 'Econometrics and Decision Theory', *Journal of Econometrics*, **95** (2), 255–283.

Clements, M. P. (2004), 'Evaluating the Bank of England density forecasts of inflation', *Economic Journal*, **114** (6), 855–877.

Clements, M. P. (2005), *Evaluating Econometric Forecasts of Economic and Financial Variables*, Basingstoke, Palgrave.

Clements, M. P. and Hendry, D. F. (1998), *Forecasting Economic Time Series*, Cambridge, Cambridge University Press.

Clements, M. P. and Hendry, D. F. (1999), *Forecasting Non-stationary Economic Time Series*, Cambridge, Mass., MIT Press.

Clements, M. P. and Hendry, D. F. (2002), 'Explaining Forecast Failure in Macroeconomics', in M. P. Clements and D. F. Hendry, eds, *A Companion to Economic Forecasting*, Blackwells Publishers, Oxford.

Clements, M. P. and Hendry, D. F. (2005), 'Guest Editors' Introduction: Information in Economic Forecasting', *Oxford Bulletin of Economics and Statistics*, **67**, 713–753.

Clements, M. P. and Hendry, D. F. (2006), 'Forecasting with Breaks', in G. Elliott, C. W. J. Granger and A. Timmermann, eds, *Handbook of Economic Forecasting*, Elsevier, Amsterdam.

Corradi, V. and Swanson, N. R. (2006), 'Predictive Density Evaluation', in G. Elliott, C. W. J. Granger and A. Timmermann, eds, *Handbook of Economic Forecasting*, Elsevier, Amsterdam.

Crafts, N. F. R. and Toniolo, G. (1996), 'Postwar Growth: an overview', in N. F. R. Crafts and G. Toniolo, eds, *Economic Growth in Europe Since 1945*, Cambridge University Press, Cambridge.

Davison, A. C. (2003), *Statistical Models*, Cambridge, Cambridge University Press.

Diebold, F., Hahn, J. and Tay, A. S. (1999), 'Multivariate density forecast evaluation and calibration in financial risk management: High-frequency returns on foreign exchange', *Review of Economics and Statistics*, **81** (4), 661–673.

Diebold, F. X., Gunther, T. A. and Tay, A. S. (1998), 'Evaluating Density Forecasts: With Applications to Financial Risk Management', *International Economic Review*, **39** (6), 863–83.

Diebold, F. X., Tay, A. S. and Wallis, K. F. (1999), 'Evaluating density forecasts of inflation: the Survey of Professional Forecasters', in R. F. Engle and H. White, eds, *Cointegrating, causality and forecasting: a festschrift in honour of Clive W. J. Granger*, Oxford University Press, Oxford.

Doornik, J. A. (2007), *Ox 5.0: an object-oriented matrix programming language*, 5th edn, London, Timberlake Press.

Doornik, J. A. and Hansen, H. (1994), 'An omnibus test for univariate and multivariate normality', mimeo, Nuffield College, Oxford.

Doornik, J. A. and Hendry, D. F. (2007), *Empirical Econometric Modelling Using PcGive 12: Volume I*, London, Timberlake Consultants Press.

Driver, R. and Wren-Lewis, S. (2005), An investigation of the implications of new trade theory for aggregate export equations using a panel of European countries, mimeo, University of Exeter and Bank of England.

Engle, R. F. and Granger, C. W. J. (1987), 'Co-integration and Error Correction: Representation, Estimation and Testing', *Econometrica*, **55** (2).

Engle, R. F. and Hendry, D. F. (2005), 'Breaks and Measurement Errors at the Forecast Origin', mimeo, University of Oxford.

Fawcett, N. W. P. (2003), *Exchange Rate Misalignment, the Balance of Payments, and the UK Economy*, Thesis, Part II Economics Tripos, University of Cambridge.

Gallant, R. (1997), *Introduction to Econometric Theory*, Princeton, Princeton University Press.

Goldstein, M. and Khan, M. S. (1985), 'Income and Price Effects in Foreign Trade', in R. R. Jones and P. B. Kenen, eds, *Handbook of International Economics*, Vol. 2, Elsevier, Amsterdam, pp. 1041–1105.

Granger, C. W. J. (1981), 'Some properties of time series data and their use in econometric model specification', *Journal of Econometrics*, **16**, 121–130.

Granger, C. W. J. (2001), 'Evaluating forecasts', in D. F. Hendry and N. R. Ericsson, eds, *Understanding Economic Forecasts*, MIT Press, Cambridge, Mass.

Granger, C. W. J. and Machina, M. J. (2006), 'Forecasting and Decision Theory', in G. Elliott, C. W. J. Granger and A. Timmermann, eds, *Handbook of Economic Forecasting*, Elsevier, Amsterdam.

Granger, C. W. J. and Pesaran, M. H. (2000*a*), 'A Decision Theoretic Approach to Forecast Evaluation', in W. S. Chan, W. K. Li and H. Tong, eds, *Statistics and Finance: An Interface*, Imperial College Press, London.

Granger, C. W. J. and Pesaran, M. H. (2000*b*), 'Economic and Statistical Measures of Forecast Accuracy', *Journal of Forecasting*, **19** (7), 537–560.

Groen, J. J. J. and Kleibergen, F. (2003), 'Likelihood-based Cointegration Analysis in Panels of Vector Error Correction Models', *Journal of Business & Economic Statistics*, **21**, 295–318.

Hall, S. G. and Mitchell, J. (2007), 'Combining density forecasts', *International Journal of Forecasting*, **23** (1), 1–13.

Hamilton, J. D. (1994), *Time Series Analysis*, Princeton, Princeton University Press.

Hendry, D. F. (1979), 'Predictive failure and econometric modelling in macro-economics: The transactions demand for money', in P. Ormerod,

ed., *Economic Modelling*, Heinemann, London, pp. 217–242. Reprinted in Hendry, D. F., *Econometrics: Alchemy or Science?* Oxford: Blackwell Publishers, 1993, and Oxford University Press, 2000.

Hendry, D. F. (1995), *Dynamic Econometrics*, Oxford, Oxford University Press.

Hendry, D. F. (2004), 'Unpredictability and the Foundations of Economic Forecasting', mimeo, University of Oxford, Oxford.

Hendry, D. F. (2006), 'Robustifying Forecasts from Equilibrium-Correction Models', *Journal of Econometrics*, **135** (1-2), 399–426.

Hendry, D. F. and Doornik, J. A. (1994), 'Modelling linear dynamic econometric systems', *Scottish Journal of Political Economy*, **41**, 1–33.

Hendry, D. F. and Ericsson, N. R. (1990), 'Modeling the demand for narrow money in the United Kingdom and the United States', *International Finance discussion paper*, **383**, Board of Governors of the Federal Reserve System, Washington.

Hendry, D. F. and Ericsson, N. R. (1991), 'Modeling the demand for narrow money in the United Kingdom and the United States', *European Economic Review*, **35**, 833–886.

Hendry, D. F. and Mizon, G. E. (1993), 'Evaluating dynamic econometric models by encompassing the VAR', in P. C. B. Phillips, ed., *Models, Methods and Applications of Econometrics*, Basil Blackwell, Oxford, pp. 272–300.

Hendry, D. F., Johansen, S. and Santos, C. (2008), 'Selecting a regression saturated by indicators', mimeo, University of Oxford.

Hlouskova, J. and Wagner, M. (2006), 'The Performance of Panel Unit Root and Stationarity Tests: Results from a large scale simulation study', *Econometric Reviews*, **25** (1), 85–116.

Hooper, P., Johnson, K. and Marquez, J. (2000), 'Trade Elasticities for the G-7 Countries', *Princeton Studies in International Economics*, **87**.

Houthakker, H. S. and Magee, S. P. (1969), 'Income and Price Elasticities in World Trade', *Review of Economics and Statistics*, **51** (2), 111–125.

Im, K. S., Pesaran, M. H. and Shin, Y. (2003), 'Testing for unit roots in heterogeneous panels', *Journal of Econometrics*, **115** (1), 53–74.

JOHANSEN, S. (1995), *Likelihood-based Inference in Cointegrated Vector Autoregressive Models*, Oxford, Oxford University Press.

JOHN, S. (1982), 'The three-parameter two-piece normal family of distributions and its fitting', *Communications in Statistics – Theory and Methods*, **11** (5), 879–85.

JÖNSSON, K. (2005), 'Cross-sectional Dependency and Size Distortion in a Small-sample Homogeneous Panel Data Unit Root Test', *Oxford Bulletin of Economics and Statistics*, **67** (3), 369–392.

KITSON, M. AND MICHIE, J. (2000), *The Political Economy of Competitiveness*, London, Routledge.

KITSON, M. AND PRIMOST, D. (2005), Corporate Responses to Macroeconomic Changes and Shocks, mimeo, University of Cambridge.

KRUGMAN, P. (1989), 'Differences in income elasticities and trends in real exchange rates', *European Economic Review*, **33** (5), 1031–1046.

KURITA, T. AND NIELSEN, B. (2004), 'Short-Run Parameter Changes in a Cointegrated Vector Autoregressive Model', mimeo, University of Oxford, Oxford.

LARSSON, R. AND LYHAGEN, J. (1999), 'Likelihood-Based Inference in Multivariate Panel Cointegration Models', *Stockholm School of Economics working papers in Economics and Finance*, **331**.

MITCHELL, J. (2005), 'A Monte-Carlo based comparison of alternative density forecast evaluation tests, with an application to the Bank of England's 'fan' chart of inflation', mimeo, NIESR.

MITCHELL, J. AND HALL, S. G. (2005), 'Evaluating, comparing and combining density forecasts using the $KLIC$ with an application to the Bank of England and NIESR "fan" charts of inflation', *Oxford Bulletin of Economics and Statistics*, **67**, 995–1033.

MPC (2004), *August Inflation Report*, London, Bank of England Monetary Policy Committee.

NICKELL, S. J. (1981), 'Biases in dynamic models with fixed effects', *Econometrica*, **49** (6), 1417–26.

NIELSEN, H. B. AND BOWDLER, C. (2006), 'Inflation adjustment in the open economy: an I(2) analysis of UK prices', *Empirical Economics*, **31** (3), 569–586.

Noceti, P., Smith, J. and Hodges, S. (2003), 'An Evaluation of Tests of Distributional Forecasts', *Journal of Forecasting*, **22** (6-7), 447–455.

OECD (2005), *The OECD STAN Database*, Paris, Economic Analysis and Statistics Division.

OECD (2006*a*), *The OECD ANBERD Database*, Paris, Economic Analysis and Statistics Division.

OECD (2006*b*), *The OECD Bilateral Trade Database*, Paris, Economic Analysis and Statistics Division.

Pagan, A. R. (2002), *Report on modelling and forecasting at the Bank of England*, London, Bank of England.

Pagan, A. R. and Ullah, A. (1999), *Nonparametric Econometrics*, Cambridge, Cambridge University Press.

Pearson, E. S., d'Agostino, R. B. and Bowman, K. O. (1977), 'Tests for departure from normality: Comparison of powers', *Biometrika*, **62** (2), 231–46.

Pedroni, P. (1995), 'Panel cointegration: asymptotic and finite sample properties of pooled time series tests with an application to the PPP hypothesis', *Indiana University Working Papers in Economics*, **95-013**.

Pedroni, P. (2000), 'Fully Modified OLS for Heterogeneous Cointegrated Panels', in B. H. Baltagi, ed., *Advances in Econometrics: Nonstationary Panels, Panel Cointegration and Dynamic Panels*, Vol. 15, Elsevier.

Pesaran, M. H. (2006), 'Estimation and Inference in Large Heterogeneous Panels with a Multifactor Error Structure', *Econometrica*, **74** (4), 967–1012.

Pesaran, M. H. (2007), 'A simple panel unit root test in the presence of cross-section dependence', *Journal of Applied Econometrics*, **22**, 265–312.

Pesaran, M. H. and Skouras, S. (2002), 'Decision-Based Methods for Forecast Evaluation', in M. P. Clements and D. F. Hendry, eds, *A Companion to Economic Forecasting*, Blackwells Publishers, Oxford.

Pesaran, M. H. and Smith, R. P. (1995), 'Estimating long-run relationships from dynamic heterogeneous panels', *Journal of Econometrics*, **68** (2), 79–113.

Pesaran, M. H. and Timmermann, A. (2004), 'How costly is it to ignore breaks when forecasting the direction of a time series?', *International Journal of Forecasting*, **20** (3), 411–425.

Pesaran, M. H. and Timmermann, A. (2005), 'Small sample properties of forecasts from autoregressive models under structural breaks', *Journal of Econometrics*, **129** (1-2), 183–217.

Pesaran, M. H., Pettenuzzo, D. and Timmermann, A. (2006), 'Forecasting Time Series Subject to Multiple Structural Breaks', *Review of Economic Studies*, **73** (4), 1057–1084.

Pesaran, M. H., Shin, Y. and Smith, R. P. (1999), 'Pooled Mean Group Estimation of Dynamic Heterogeneous Panels', *Journal of the American Statistical Association*, **94** (446), 621–634.

Phillips, P. C. B. and Moon, H. R. (1999), 'Linear regression limit theory for nonstationary panel data', *Econometrica*, **67** (5), 1057–1111.

Phillips, P. C. B. and Sul, D. (2003), 'Dynamic panel estimation and homogeneity testing under cross section dependence', *Econometrics Journal*, **6** (1), 217–259.

Poulizac, D., Weale, M. and Young, G. (1996), 'The Performance of National Institute Economic Forecasts', *National Institute Economic Review*, pp. 55–62.

Rao, C. R. (1973), *Linear Statistical Inference and Its Applications*, 2nd edn, New York, John Wiley and Sons.

Redding, S. (2002), 'Path Dependence, Endogenous Innovation, and Growth', *International Economic Review*, **43** (4), 1215–1248.

Rosenblatt, M. (1952), 'Remarks on a multivariate transformation', *Annals of Mathematical Statistics*, **23** (4), 470–2.

Sarno, L. and Valente, G. (2004), 'Comparing the accuracy of density forecasts from competing models', *Journal of Forecasting*, **23** (8), 541–557.

Silverman, B. W. (1986), *Density estimation for statistics and data analysis*, London, Chapman and Hall.

Smith, R. P. and Fuertes, A.-M. (2007), 'Panel Time-Series', Mimeo, Department of Economics, Birkbek College, London and City University, London.

Tay, A. S. and Wallis, K. T. (2002), 'Density Forecasting: A Survey', in M. P. Clements and D. F. Hendry, eds, *A Companion to Economic Forecasting*, Blackwells Publishers, Oxford.

Thompson, S. (2004), 'Identifying term structure volatility from the LIBOR swap curve', mimeo, University of Harvard.

Timmermann, A. (2000), 'Density Forecasting in Economics and Finance', *Journal of Forecasting*, **19** (4), 231–234.

Ullah, A. (1996), 'Entropy, divergence and distance measures with econometric applications', *Journal of Statistical Planning and Inference*, **49** (2), 137–162.

Wagner, M. and Hlouskova, J. (2007), 'The Performance of Panel Cointegration Methods: Results from a Large Scale Simulation Study', *Institute for Advanced Studies Working Papers*, **210**, Forthcoming in *Econometric Reviews*.

Wallis, K. F. (1999), 'Asymmetric density forecasts of inflation and the Bank of England's fan chart', *National Institute Economic Review*, **39** (167), 106–112.

Wallis, K. F. (2004), 'An Assessment of Bank of England and National Institute Inflation Forecast Uncertainties', *National Institute Economic Review*, **47** (189), 64–71.