EMPIRICAL MODELLING AND MODEL SELECTION FOR FORECASTING INFLATION IN A NON-STATIONARY WORLD

EMPIRICAL MODELLING AND MODEL SELECTION FOR FORECASTING INFLATION IN A NON-STATIONARY WORLD

Jennifer L. Castle

Nuffield College
University of Oxford

This thesis is submitted
for the degree of Doctor of Philosophy
in the subject of Economics
at the University of Oxford

Hilary Term 2006

Contents

List of Figures

List of Tables

Preface

Borrowing from David Hendry's forecasting analogy, writing a thesis is akin to embarking on a long journey in which the destination is unknown. The thesis has consisted of many wrong turns, diversions and traffic jams. Regardless of the amount of route planning involved, getting lost along the way has been an inevitability. However, at many junctions, opting for the more scenic of the potential routes has sustained my enthusiasm and enjoyment of the entire journey, regardless of the destination. Of course, the journey was not undertaken alone, and I owe a tremendous debt to a great many people who have accompanied me along the way.

My greatest and most sincere thanks go to my supervisor, David Hendry, who has been my constant source of inspiration throughout the journey. Without his awe-inspiring intellectual capacity, his inexhaustible supply of ideas, his perpetual encouragement, and his never-failing optimism, this thesis would not have reached fruition. My enthusiasm and passion for econometric research is a direct result of working with David, and it has been an honour and a privilege to hold the positions of both David's student and research officer. I can only hope that I shall be fortunate enough to continue to work with David in the future.

Enjoyment of the thesis was particularly enhanced by my fellow PhD students at Oxford University. The team began with Sonja Keller, Carlos Santos, and myself, and gradually expanded to include James Reade and Nicholas Fawcett. It has been a joy to work in such a supportive team, and I have greatly benefited from intellectual discussions in the many Thursday morning meetings and endless coffee breaks.

The Department of Economics has provided an excellent, stimulating environment that has been invaluable to my research. I would like to thank Sujit Kapadia and Takamitsu Kurita for their advice and friendship. My thanks extend to all the staff at the department, whose friendly faces and willingness to help ensured a pleasurable working atmosphere.

The Nuffield College community has played a fundamental role in both my intellectual and social enjoyment of Oxford. My particular thanks go to my college supervisor, Neil Shephard, who has been extremely supportive throughout my time at Nuffield. I have also benefited immensely from the help and advice of Bent Nielsen, Jurgen Doornik, John Muellbauer, Christopher Bowdler, Roland Meeks and Clive Bowsher. Furthermore, both the staff and students of Nuffield College have ensured that my Nuffield affiliation has been delightful, and my particular thanks go to Maureen Baker, Margaret Yee, Jenny Haydock, Jane Green, Tehmina Khan, Alberto Behar, David Gill, Victoria Prowse and Julia Giese.

It is a great pleasure to thank the Bank of England for providing me with the opportunity to experience econometrics 'in the real world' over successive summers. Colin Ellis receives my earnest gratitude for bestowing me with a co-authorship, and further thanks go to Spencer Dale, Andy Haldane, Simon Hayes, Alex Brazier, Alisdair Scott, Clare Lomardelli, Iain De Weymarn, Gregory Thwaites and Orla May, who all gave me excellent supervision, advice or data.

I am indebted to numerous academics who have provided help or advice along the way, including Kirstin Hubrich, Neil Ericsson, Patrick Royston, Richard Gerlach, Anders

Rahbek, Gunnar Bårdsen, Adrian Pagan, Karim Abadir and Charles Bos. I am most grateful to Hans-Martin Krolzig, both for his Monte Carlo code and for his responses to my numerous questions regarding PcGets. I thank Teodosio Perez-Amaral, Giampiero Gallo, Halbert White and Max Marinucci for the RETINA code, and my particular thanks extend to Christian Brownlees, who was always willing to help with my queries regarding RETINA.

I would like to thank participants of the Gorman Workshop, Nuffield College, February 2004; the Department of Economics Seminar, Oxford University, February 2005; the Econometric Study Group Conference, Bristol, July 2005; the 3rd OxMetrics Users Conference, Cass Business School, London, August 2005; the workshop on Nonlinear and Nonstationary Time Series, Kaiserslautern, September 2005; and the IASC World Conference on Computational Statistics and Data Analysis, Cyprus, October 2005, for insightful comments on my papers. Furthermore, I would like to thank Erasmus University, Rotterdam, who hosted the Journal of Applied Econometrics Annual Lectures, May 2004; Heidelberg University, who hosted the NBER/NSF Time Series Conference, September 2005; and the European Central Bank, Frankfurt, for their hospitality.

It is a great pleasure to thank the Economic and Social Research Council, who provided financial support for my thesis, both in the form of a Postgraduate Studentship Award and as a Research Officer under grant RES 051 270035.

Without my friends outside of economics, I may not have survived the journey with my sanity in tact. My sincerest gratitude goes to my long-term flat mate, Sarah Butt, for always lending an ear throughout the ups and downs of the thesis. My closest friends have been an invaluable source of support throughout the thesis, and my heartfelt thanks go to Jayne Almond, Lucy Bamford, Saskia Ouwendijk and Samantha Mudge.

Of course, the greatest thanks go to my family, without whose support I could not have completed the thesis. They deserve gold medals for listening to my numerous, and often incomprehensible, explanations of econometric methodology!

I started this journey with my parents, Bob and Wendy, who gave me never-failing moral support. The journey must end without my father by my side to witness the finished product. I dedicate this thesis to him, whose infectious enthusiasm and joie de vivre, along with his constant encouragement and unconditional belief in me, provided me with the confidence and conviction to achieve my goals. He gave me the courage to commence this thesis and I only wish he could witness its completion. This thesis is in memory of Bob Castle.

Jennifer L. Castle
Nuffield College
March 2006

Abstract

Selection and forecasting are integral to econometric modelling and this thesis addresses both issues, with an application to UK inflation. Two automatic model selection algorithms, PcGets and RETINA, are evaluated on time-series and cross-section data. PcGets aims to select an undominated, congruent model of the phenomena of interest, whereas RETINA selects a parsimonious set of regressors that have predictive ability. Monte Carlo simulation results assess the null and non-null rejection frequencies of the algorithms in the presence of nonlinear functions. Both algorithms have the same properties as those for linear models under orthogonality, but collinearity increases null rejection frequencies and reduces non-null rejection frequencies. Simple operational rules that 'double de-mean' all functions are proposed to mitigate that problem.

A nonlinear model selection strategy is proposed, that commences with a new test for nonlinearity, specifies the general model using polynomial functions as approximations, and undertakes a general-to-specific reduction using a multi-stage procedure. Nonlinearity poses a number of problems, including collinearity generated by nonlinear transformations, extreme observations leading to non-normal (fat-tailed) distributions, and often more variables than observations from general expansions approximating the nonlinearity, yet one must avoid excess retention of irrelevant variables. Solutions to all of these problems are proposed. A successful algorithm requires the synthesis of all of these developments to be implemented, as exclusion of one component of the algorithm can lead to severely erroneous conclusions.

A model of inflation is built in which many determinants of inflation play a role in its explanation. The single cause explanation of inflation is refuted, along with a generic business cycle explanation. As forecast failure is prevalent, with naive devices often outperforming econometric models, a forecast competition is undertaken for UK annual and quarterly inflation, in which equilibrium correction models are compared to various forecasting rules. Robust forecasting devices prove useful in forecasting macroeconomic time-series, and they often outperform econometric models, both when there are structural breaks in the data and when the underlying process appears to be stable but with breaks in the explanatory variables. Increasing the information set does lead to improvements in forecasting performance suggesting that disaggregation can yield benefits. It is observed that much of the forecast error in the structural models is driven by the deterministic terms. Breaks in the mean of the cointegrating vector or the growth rate of the system will cause forecast 'failure', and results show how sensitive forecasts are to errors in these terms.

JEL classification: C22; C32; C51; C53; C87.

Keywords: Econometric methodology; Model selection; Nonlinear models; Nonlinearity testing; Collinearity; Inflation modelling; Robust forecasting; Inflation forecasting.

Statement of Publication and Word Count

Parts of the thesis have been published or have been accepted for publication. Chapter 2 of the thesis, entitled 'Evaluating PcGets and RETINA as Automatic Model Selection Algorithms', has been published in the *Oxford Bulletin of Economics and Statistics*, Volume **67**, pp. 837-880. Parts of Chapter 4 and Chapter 5 have been accepted to appear in Rapach and Wohar (eds), *Forecasting in the Presence of Structural Breaks and Model Uncertainty*, which is part of the new Elsevier Series on the Frontiers of Economics and Globalization, under the title 'Forecasting UK Inflation: Empirical Evidence on Robust Forecasting Devices' with David F. Hendry.

The papers that comprise the thesis have been presented at various seminars and conferences. Chapter 2, 'Evaluating PcGets and RETINA as Automatic Model Selection Algorithms', has been presented at the Department of Economics Seminar, Oxford University, February 2005. Chapter 3, 'Extending the Boundaries of PcGets: Nonlinear Models', has been presented at the Econometrics Seminar, Nuffield College, June 2005; the Econometric Study Group Conference, Bristol, July 2005; the 3rd OxMetrics Users Conference, Cass Business School, London, August 2005; the workshop on Nonlinear and Nonstationary Time Series, Kaiserslautern, September 2005, which was a satellite workshop of the NBER/NSF Time Series Conference in Heidelberg; and the IASC World Conference on Computational Statistics and Data Analysis, Cyprus, October 2005. A previous version of Chapter 4, titled 'Measuring Excess Demand and its Impact on Inflation', was presented at the Gorman Workshop, Nuffield College, February 2004. Finally, Chapter 5, 'Forecasting UK Inflation: Empirical Evidence on Robust Forecasting Devices', has been presented at the Annual Conference of the Scottish Economic Society, April 2006.

Several applications were used to produce the thesis, figures and calculations: the text was set using LaTeX; *Ox* was used for computations and simulations (Doornik, 2001); *PcNaive* was also used for simulations (Doornik and Hendry, 2001b); *GiveWin* was used for graphics (Doornik and Hendry, 1996); *PcGive* was the main econometrics package used (Doornik and Hendry, 2001a, 2001c); *STAMP* was used for the structural time series models (Koopman, Harvey, Doornik and Shephard, 1995); the model selection algorithms used include *PcGets* (Hendry and Krolzig, 2001) and *RETINA* (Marinucci, 2005). A Matlab version of the *RETINA* code was utilised, which can be downloaded from www.ds.unifi.it/ctb/projects.html. A new windows version of *RETINA* is available at http://personales.ya.com/max_mar/retina_v0b_p.exe. Full citations for the applications used are provided in the text.

WORD COUNT

The body of this thesis comprises 232 pages (including references), with a typical page of text containing around 300 words, making the document approximately 69,600 words long.

Nomenclature

Model selection terminology

DGP	data generating process
LDGP	local data generating process
GUM	general unrestricted model
ADL	autoregressive distributed lag
LDV	lagged dependent variable
Gets	general-to-specific
Congruence	the model matches the evidence in all measured respects (see Hendry, 1995, p.365)
RSS	residual sum of squares: $\sum_{t=1}^{T} (y_t - \widehat{y}_t)^2$
R^2	squared multiple correlation: $1 - \frac{\sum \widehat{\epsilon}_t^2}{\sum (y_t - \bar{y})^2}$, where $\sum \widehat{\epsilon}_t^2 = \sum (y_t - \widehat{y}_t)^2$
$\overline{\mathsf{R}^2}$	$1 - \frac{T-1}{T-k}\left(1 - \mathsf{R}^2\right)$
IC	information criteria
AIC	$\ln \widetilde{\sigma}_k^2 + \frac{2k}{T}$ (see Akaike, 1973)
SIC	$\ln \widetilde{\sigma}_k^2 + c\frac{k \ln T}{T}$ (see Schwarz, 1978)
HQ	$\ln \widetilde{\sigma}_k^2 + \frac{2k \ln(\ln T)}{T}$ (see Hannan and Quinn, 1979)
	where $\widetilde{\sigma}_k^2 = T^{-1} \sum_{t=1}^{T} (y_t - \widehat{y}_t)^2$

Empirical modelling terminology

PF	production function
HP	Hodrick-Prescott filter (see Hodrick and Prescott, 1997)
UC	unobserved components model
SSF	state space form
TFP	total factor productivity
SR	Solow residual
NAIRU	non-accelerating inflation rate of unemployment
PPP	purchasing power parity
PC	principal components
VAR	vector autoregression
VEqCM	vector equilibrium correction mechanism
EqCM	equilibrium correction mechanism (single-equation)
ARMA	autoregressive moving average
VARMA	vector autoregressive moving average
ARIMA	autoregressive integrated moving average
ARFIMA	fractionally-integrated autoregressive moving average
STR	smooth transition regression
STR-D	smooth transition regression - deviation model

LSTAR	logistic smooth transition autoregression
TAR	threshold autoregression
SETAR	self-exciting threshold autoregression
EAR	exponential autoregression
ARCH	autoregressive conditional heteroskedasticity
GARCH	generalized autoregressive conditional heteroskedasticity
RCA	random coefficient autoregression
ANN	artificial neural network
MSM	Markov-switching model
DSGE	dynamic stochastic general equilibrium

Diagnostic testing terminology

$\mathsf{F_{ar}}(\mathsf{k},.)$	k^{th}-order serial correlation (see Harvey, 1990)
$\mathsf{F_{arch}}(\mathsf{k},.)$	k^{th}-order autoregressive conditional heteroskedasticity (see Engle, 1982)
$\chi^2_{\mathsf{nd}}(2)$	Doornik-Hansen test for normality (see Doornik and Hansen, 1994)
$\mathsf{F_{het}}(.)$	heteroskedasticity test using squares (see White, 1980)
$\mathsf{F^{np}_{het}}(.)$	non-parametric test for heteroskedasticity (see Koopman *et al.*, 1995, p.183)
$\mathsf{F_{reset}}(.)$	regression specification test (see Ramsey, 1969)
$\mathsf{F_{Chow}}(.)$	break-point and forecast Chow tests (see Chow, 1960)
$\chi^2_{\mathsf{bl}}(\mathsf{k})$	portmanteau Box-Ljung Q-statistic for k^{th}-order serial dependence (see Ljung and Box, 1978)
DW	Durbin-Watson test for autocorrelated residuals (see Durbin and Watson, 1951)
$\mathsf{SC_r}(\mathsf{k})$	residual autocorrelation at lag k (see Koopman *et al.*, 1995, p.173)
ADF	augmented Dickey-Fuller test (see Dickey and Fuller, 1981)
HACSE	heteroskedasticity and autocorrelation consistent standard errors (see Andrews, 1991)

Forecast terminology

ΔVEqCM	differenced vector equilibrium correction mechanism
ΔVEqCM$_\beta$	differenced vector equilibrium correction mechanism excluding I(-1) terms
ΔEqCM	differenced equilibrium correction mechanism (single-equation)
ΔEqCM$_\beta$	differenced equilibrium correction mechanism excluding I(-1) terms
DV	differenced vector autoregression
DDV	double differenced vector autoregression
SMD	smoothed difference model: $\Delta x_t = \frac{1}{4}\Delta_4 x_{t-1}$
RW	random walk
Pool	pooled forecast
EWMA	exponentially weighted moving average
ME	mean error: $\frac{1}{H}\sum_{t=T+1}^{T+H}(y_t - \widehat{y}_t)$
MAE	mean absolute error: $\frac{1}{H}\sum_{t=T+1}^{T+H}\lvert y_t - \widehat{y}_t\rvert$
MAPE	mean absolute percentage error: $\frac{100}{H}\sum_{t=T+1}^{T+H}\left\lvert\frac{y_t-\widehat{y}_t}{y_t}\right\rvert$
MSFE	mean square forecast error: $\frac{1}{H}\sum_{t=T+1}^{T+H}(y_t - \widehat{y}_t)^2$
RMSFE	root mean square forecast error: $\sqrt{\mathsf{MSFE}}$

CMSFE	cross-validated mean square forecast error
GFESM	generalised forecast error second moment
MGN	Morgan-Granger-Newbold test (see Granger and Newbold, 1986)

Econometric notation

OLS	ordinary least squares
IV	instrumental variables
L	lag operator. $Lx_t = x_{t-1}$
Δ_j^i	difference operator, defined as $\left(1 - L^j\right)^i$. Subscripts are omitted for first differences but are indicated for fourth differences, denoting the annual change
I_{date}	indicator variable. $I_{date} = 1_{(t=date)}\ \forall T$
$D_{date_{i,j}}$	dummy variable. $D_{date_{i,j}} = \begin{cases} 1_{(t=date_i)} \\ -1_{(t=date_j)} \end{cases} \forall T$

Graphical notation

Figure panels are lettered notionally as $\begin{bmatrix} a & b \\ c & d \end{bmatrix}$

EMPIRICAL MODELLING AND MODEL SELECTION FOR FORECASTING INFLATION IN A NON-STATIONARY WORLD

Introduction

Forecasting is integral to economic policy and yet macroeconomic forecasting has experienced many episodes of serious predictive failure. It is evident that the economy is non-stationary and subject to numerous structural breaks, data are contaminated with measurement errors and models are mis-specified, but empirical econometric models tend not to account for these problems. Often, forecasting models are developed assuming the underlying data generating process (DGP) is stationary and the DGP coincides with the econometric model. While forecasting theory is well established for such a scenario (see Granger and Newbold, 1986) it is of little practical use in applied work. A new theoretical framework for forecasting has been developed by Clements and Hendry (1998b, 1999), which allows for both mis-specified models and non-stationarities including structural change. This thesis aims to address the dual issues of model specification and non-stationarity in a pragmatic way, assessing their generic implications and providing practical solutions to the forecasting conundrum.

The assessment of forecasting models requires a constructive and germane application, and we focus on UK inflation. Inflation forecasting is fundamental to macroeconomic policy, given that it is the sole target of monetary policy in the UK, and the move towards an inflation targeting regime has put pressure on inflation forecasting models to deliver timely, unbiased and efficient forecasts of future inflation. However, forecast failure is prevalent, with naive devices often outperforming the dominant congruent in-sample model in forecasting competitions.[1] This is because structural change is endemic in any economy. Even during relatively quiescent periods, unanticipated shifts occur frequently and are a major source of forecast failure. We investigate the use of robust

[1]The dominant congruent in-sample model refers to a model that is coherent with the available evidence, assessed by a range of mis-specification tests, and is dominant in the sense that it encompasses other model specifications, (see Hendry, 1995, p.365). The dominant congruent in-sample model can be viewed as the 'best' in-sample econometric model built given the current level of knowledge and empirical evidence.

forecasting devices to improve forecasts of UK inflation.

The theme that underpins all chapters of the thesis is model selection. The development of any econometric model requires model selection rules of some form, because economic processes are extremely complex and the underlying DGP is unknown. Model selection is difficult because all tests have interdependent distributions that are altered by every modelling decision. Furthermore, different model selection rules may be required for in-sample modelling and for forecasting. Our focus is on general-to-specific (*Gets*) model selection, which has a strong theoretical motivation in the theory of reduction. The automatic model selection algorithm, PcGets, is utilised in all four substantive chapters, demonstrating our belief that automatic *Gets* selection tools are integral to modelling and forecasting in a non-stationary world.

The thesis is broadly divided into two related parts that each address modelling and forecasting when the DGP is evolutionary, non-stationary, and unknown to the econometrician. These include specification, in which the model may not coincide with the DGP, and non-stationarity, in which the DGP is not constant, but is instead evolving over time and is subject to intermittent structural breaks. We further partition both parts of the thesis into two subcategories, which include model selection and model evaluation. Table 1 outlines the underlying structure of the thesis. In [I.a.] we consider techniques to select econometric models for forecasting purposes and examine the performance of two automatic econometric model selection algorithms, PcGets and RETINA, in selecting forecasting models. In [I.b.] we outline a general-to-specific selection algorithm for selecting nonlinear models, developing a test of functional form and proposing strategies to overcome collinearity, non-normality, and selection with a large number of regressors. Section [II.a.] builds a model of UK inflation within the linear *Gets* framework, and [II.b.] examines the use of robust forecasting devices in forecasting UK inflation.

The structure of the thesis is as follows. Chapter 2 examines the selection of forecasting models using automatic econometric model selection techniques. Despite the controversy surrounding many model selection strategies, and the multitude of diverse approaches in the literature, automatic model selection procedures are fast becoming one

I.		**Specification**
	a.	Model selection
	b.	Model evaluation
II.		**Non-stationarity**
	a.	Model selection
	b.	Model evaluation

Table 1: Thesis structure

of the dominant methods of model building. To date, the focus has primarily concentrated on developing in-sample models, and this chapter considers the use of automatic algorithms when building forecasting models. We compare PcGets and RETINA, assessing their selection properties and forecasting performance on both cross-section and time-series data, as well as Monte Carlo evidence. PcGets is based on a *Gets* search strategy, starting with a general model capturing the underlying characteristics of the data and testing downwards, ensuring validity of the reductions at each stage to result in a congruent, parsimonious, undominated model. RETINA differs from PcGets in that the *Gets* methodology is not its main tenet. The program generates many transformations capturing both nonlinearities and interaction terms, and then adds variables into the model depending on a given criterion. RETINA aims to identify a set of variables that are likely to be relevant for predicting out-of-sample.

The RETINA selection algorithm raises the question of model selection when there are more variables than observations. The set of nonlinear functions generated can result in a very large set of potential regressors over which to search, and it is not uncommon for there to be more potential regressors than observations, particularly with sample sizes available in macroeconomic time-series data. We extend work by Hendry and Krolzig (2003a) to establish the small sample properties of selection with more variables than observations. A substantive, but mainly hidden, component of the thesis has been the development of computer code to undertake many of the calculations, embed RETINA in a simulation algorithm, and undertake many new simulation studies.

As RETINA selects models from a set of nonlinear regressors and PcGets can easily be generalised to select corresponding nonlinear models, as well as the conjecture that the DGP is indeed nonlinear (see Granger and Teräsvirta, 1993, p.1), Chapter 3 considers

selection of nonlinear models. A comprehensive selection strategy is developed, that commences with a test of functional form, generates a nonlinear unrestricted model, and then applies a multi-stage procedure to select, using tighter significance levels for the nonlinear functions.

A test of functional form is developed that is similar to the test for heteroskedasticity proposed by White (1980), where the heteroskedastic-consistent covariance matrix estimator will differ from the conventional formula when the squares and cross-products of the regressors would be significant if added to the model. There are three main drawbacks of a test such as White's: first, its high dimensionality; secondly, the potentially high level of collinearity between products of regressors; and third, the possibility that the second derivative is not the source of the important departure from linearity, which may depend on asymmetry or skewness and be better reflected in the third derivative. To rectify these potential drawbacks, our test forms a composite function of all product terms of the regressors, based on weights given by the eigenvalues of the variance-covariance matrix. Thus, for fixed regressors the test is an exact F-test with n degrees of freedom (where n is the number of regressors) on the standardised, mutually orthogonal combinations of squares and cross-products of the original data matrix, solving the problems of high dimensionality and collinearity. The test will not have power if the second derivative is zero but higher orders are non-zero, but in this case, cubic terms of the composite function could be included as well, yielding an F-test with $2n$ degrees of freedom. The test is a useful component of the selection strategy for nonlinear models, and is easily embedded into an automatic *Gets* framework.

If evidence of nonlinearity is found, but the functional form is unknown, an approximation is needed. We consider polynomial functions as they approximate a range of DGPs, including smooth transition models, which nest aspects of Markov-switching models, bilinear models and neural networks. Rigorous mis-specification testing is undertaken to ensure the local data generating process (LDGP) is nested within the polynomial approximating class. As the selected nonlinear model should nest its linear counterpart, the reduction to a linear model can be tested.

The chapter also addresses a number of concerns with selecting nonlinear models. First, collinearity may be problematic because a high degree of correlation can be generated between the original linear variable and the corresponding nonlinear transformation. The information content of an extra highly collinear variable is very small but it disrupts existing information attribution. Simple operational rules are developed to transform the model to a more orthogonal form prior to selection. Second, we address the problem of extreme observations. Fat-tailed distributions are problematic for any model selection procedure because the assumption of normality is often inbuilt for the critical values used by model selection procedures. The problem is accentuated when nonlinear functions 'line up' with extreme observations, leading to nonlinear functions being retained too often. We propose the solution of removing extreme observations using indicator saturation techniques, developed by Hendry, Johansen and Santos (2004), to ensure near normality for inference. Finally, we consider the significance levels used for selecting models. While it is commonly thought that parsimony is crucial for successful forecasting models, we reject this claim, asserting instead that robustness is the essential criterion. Hendry and Hubrich (2006) provide evidence that large models are preferable for forecasting, with variables contributing to the forecast if the non-centrality of the t^2-statistic from the DGP is greater than unity. On the other hand, retention of irrelevant nonlinear functions is problematic for forecasting, leading to a loss in robustness. We propose the use of a 'super-conservative' strategy for the selection of nonlinear functions while retaining standard significance levels for the linear functions, and outline a method for implementation of that strategy.

Chapters 4 and 5 focus on section II, examining issues of non-stationarity using an inflation forecasting application. Non-stationarity not only refers to non-constant moments, but also to intermittent shifts and structural breaks. Chapter 4 builds a model of inflation using the single-equation *Gets* framework of PcGets. Measures of excess demand for both goods and services and labour are developed. As the output gap is a latent variable, it is difficult to obtain an accurate estimate of excess demand pressures. Our preference is to develop a measure based on the production approach,

as this has the most rigorous theoretical foundation. Concerns over measurement errors led to the development of a dynamic production function measure, in which total factor productivity (TFP) is modelled as a random walk with drift. However, the resulting trend estimates are too volatile, and hence a Solow residual framework is used, in which TFP is captured by a segmented linear trend. A congruent model of quarterly UK inflation is developed, based on excess demands from all sectors of the economy captured by equilibrium correction terms. Most extant theories of inflation play a role in its determination, including unit labour costs, import prices, producer prices, exchange rates, foreign prices and excess demands for both goods and labour. Furthermore, evidence for a generic 'business cycle' factor driving inflation is limited: a business cycle component derived using principal components analysis does not negate the individual variables included, and very few principal components are retained. The lack of interpretability and non-robustness to changes in the information set limit the use of a broad cyclical factor.

Chapter 5 examines the forecast performance of various inflation models. A vector equilibrium correction model of inflation is built based on various input prices and conditioning on excess demand, to generate dynamic forecasts. The forecast performance of both econometric models, including the single-equation model developed in Chapter 4 and the multi-equation model developed in this chapter, are examined against various robust forecasting devices, including two alternative differenced equilibrium correction models, a differenced VAR based on a 5−year rolling average growth rate, a random walk, an autoregressive model, a smoothed difference model, and a pooled forecast. Robust forecasting devices do prove useful in forecasting macroeconomic time-series, and they often outperform the dominant congruent in-sample models, both when there are structural breaks in the data and when the underlying process appears to be stationary but with breaks in the explanatory variables. We observe that much of the forecast error in the structural models is driven by the deterministic terms. Breaks in the mean of the cointegrating vector or the growth rate of the system will cause forecast 'failure' and results show how sensitive forecasts are to errors in these terms.

The chapter also investigates an empirical application of the theory of predictability (see Clements and Hendry, 2005a). Predictability is relative to the information set that is used, and forecasting from a reduced, but proper, information set should produce less accurate but unbiased predictions. Temporal disaggregation cannot lower predictability, and so as lower frequency data is a subset of higher frequency data, more accurate predictions of annual inflation should be obtained using quarterly data as opposed to annual data, although both forecasts should be unbiased. We find that increasing the information set by disaggregation within the time dimension can yield benefits. Increasing the frequency of data should lead to a faster reaction to observed structural breaks, implying that a shift towards robust forecasting methods could be implemented before substantial and prolonged forecast failure is realised. Conversely, measurement errors may have a more substantial impact at the forecast origin for higher frequency data. We have abstracted from measurement error in this thesis but observe that it is an important component of econometric modelling and forecasting, and we leave the topic for future research.

The Epilogue concludes the thesis, drawing together the research from the substantive chapters to provide a coherent and feasible approach to building econometric models for forecasting. We have shown that automatic model selection algorithms are integral to building econometric models, and have demonstrated the viability of generalising PcGets to select nonlinear models. The discipline is rapidly focusing on the use of nonlinear models for forecasting, and we have addressed concerns that may arise when forecasting with nonlinear models. Furthermore, we assessed the evidence for using robust forecasting devices and found that 'non-optimal' models may well deliver forecasting gains. This thesis aims to make some progress towards explaining the prevalence of forecast failures, and to provide solutions that may help to overcome these problems.

Chapter 1
A Survey of the Literature

1.1 Introduction

This chapter provides a literature review of the fields of model selection and forecasting. There is a substantial literature concerned with the selection of in-sample models, and an equivalently exhaustive literature regarding forecasting from a postulated forecasting model. The literature examining the selection of a model for the purpose of forecasting is perhaps more sparse, but has a long history stemming from model selection based on information criteria.[1] One potential explanation is that in-sample selection criteria is often thought to be sufficient for forecasting models. This is due to the prevalent assumption that the optimal forecasting model is the dominant in-sample model. Conventional forecast theory, commencing from the seminal works of Box and Jenkins (1970) and Klein (1971), is based on this premise. Granger and Newbold (1986) prove that under assumptions of stationarity and the correct model (i.e. the model and DGP coincide), the best in-sample model is the best out-of-sample model, hence delivering the lowest mean square forecast error (MSFE) matrix.

The assumptions required for conventional forecasting theory to hold are restrictive and unrealistic. The world is non-stationary, experiencing both stochastic unit roots and unanticipated structural breaks. Furthermore, the underlying DGP is extremely complicated and our understanding of the world, and subsequent modelling techniques, is incomplete at best, delivering simplified approximations of a complicated, ever-changing process. The Clements and Hendry (1998b, 1999) theory of forecasting relaxes the assumptions of conventional theory, allowing for a non-stationary and evolving world in which the DGP and econometric model differ. In this paradigm, non-causal models can

[1] More recent research includes Phillips (1994, 1995, 1996, 2003), who proposes a selection algorithm that uses posterior information criteria to select forecasting models, and Swanson and White (1997a, 1997b), who build forecasting models using artificial neural networks. Clements and Hendry (2005a) and Hendry and Hubrich (2006) establish optimal selection criteria for selecting forecasting models based on the theory of predictability.

outperform causal models (see Hendry, 1997a), implying that methods of model selection may well differ if the purpose is forecasting as opposed to in-sample analysis.

The thesis is embedded within a progressive research strategy, see Hendry (1995, p.550) and Hendry (2000b, p.177). This approach does not require complete information prior to developing empirical models, as is the case for models built on theory alone (which Keynes, 1939, 1940, argued was a precondition for any inference from empirical models). The approach gradually accumulates knowledge, conditional on the stock of information previously accumulated. Hence, we build on both the general-to-specific approach to model selection, and the inflation modelling approach that allows for many explanations of inflation, captured in the papers of, *inter alia*, De-Brouwer and Ericsson (1998), Hendry (2001) and Bårdsen, Jansen and Nymoen (2003b). Furthermore, we use the techniques developed in Clements and Hendry (1999), Hendry and Clements (2000), Hendry (2005b) and Hendry and Mizon (2005) to build alternative forecasting models. The literature review aims to identify the stock of knowledge on which the thesis builds.

To outline the structure of the chapter, Figure 1.1 demonstrates the linkages between the broad topics that are reviewed, each relating to a substantive chapter of the thesis. We approach the literature from the perspective of model selection; the underlying theme of the thesis. Section 1.2 discusses the current literature on model selection, briefly addressing the criticisms associated with data mining, with section 1.3 focusing on automatic model selection procedures. Section 1.4 examines nonlinear models, reviewing the vast array of both nonlinear models and tests for nonlinearity proposed in the literature. The interactions between automatic model selection and nonlinear models are addressed in Chapter 3. Section 1.5 looks at the literature on inflation modelling, focusing on the equilibrium correction class in which all theories of inflation can potentially play a role. The linkages between automatic model selection techniques and inflation models are demonstrated in Chapter 4. Finally, section 1.6 addresses the current standpoint of the forecasting literature, highlighting the problem of structural breaks and motivating the use of robust forecasting devices used in Chapter 5.

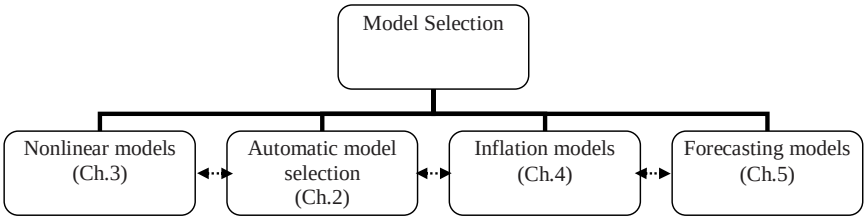

Figure 1.1: Literature survey overview

1.2 Model selection

Model selection is ubiquitous. As the underlying DGP is unknown, some method of approximating that entity is required. The techniques used can broadly be divided into two non-distinct catagories; those that are built using economic theory and those that are based on empirical data, see Hendry (1980). Hendry and Mizon (2000) provide a critique of models developed to test economic theories, supporting our empirical strategy of focusing on data-based model selection.

Data mining has courted much controversy in the field of economics; its connotations of distorted statistical inference resulting from a prejudiced search are pervasive in the literature, and yet there is an increase in data mining, partly due to the ever increasing size of databases and the availability of computing power and algorithms to analyse them, and partly due to the realisation that building models on the basis of theory alone is insufficient to develop models for analysis, policy and forecasting. The data mining controversy is not new: Frisch (1934) advocated methods to deal with data mining and Haavelmo (1944) also discusses the problem. Concerns over data mining came to the fore again with Leamer (1978), supported by the empirical evidence of Lovell (1983) who found low success rates for selecting a small model within a large database. Furthermore, there were concerns as to the apparent coefficient bias when selecting using significance tests and the claimed under-estimation of coefficient standard errors, see Campos, Ericsson and Hendry (2005). However, Gilbert (1986) argues that stringent and critical model evaluation can avoid data mining, and Hendry (1995, ch.15), Campos and Ericsson (1999), and Sargan (2001), demonstrate that data mining is not insurmountable. Constructive data mining is a procedure that yields a high probability of locating the DGP while 'letting the data speak for itself', and hence is a useful model selection procedure.

There have been many *ad hoc* selection procedures based on model fit, including R^2 and $\overline{R^2}$ criteria (although these measures do not account for the costs associated with selecting an incorrect model); Mallows (1973) C_p criterion, based on mean square prediction error; and Amemiya (1980) prediction criterion, based on the unconditional mean square prediction error. However, Hendry and Krolzig (2005) observe that a model should not be selected on the basis of fit, but instead, selecting the minimal congruent encompassing model will result in the best fit. Information criteria penalise the log-likelihood function by a function of the number of parameters to induce a parsimonious explanation of the data phenomena, and are well established in the model selection literature (see, *inter alia*, Akaike, 1973, Schwarz, 1978, Hannan and Quinn, 1979, Chow, 1981b, and Phillips, 1994, 1995). Both the Hannan-Quinn and Schwarz criterion deliver consistent model selection, see, e.g., Campos, Hendry and Krolzig (2003), but an information criterion is not a sufficient principle upon which to select models as it does not ensure congruency, and so a mis-specified model could be selected (see Bontemps and Mizon, 2003).

Shrinkage techniques are popular in the statistics literature, stemming from the work of Stein (1956) who showed that maximum likelihood and least squares principles are unsuitable for handling many parameters. The technique uses the data to determine the compromise between bias and variance, which results in a transformation of the parameter vector such that all the information in the GUM is retained, but a smooth discount weight is applied to the coefficient estimates, see James and Stein (1961) who develop the first estimator, and Yancey and Judge (1976) and Judge and Bock (1978). While this approach has been proposed as a solution to the pre-test problem, it does deliver biased estimates. Furthermore, the technique is not progressive, in the sense of knowledge accumulating about the process being modelled, because the decision rule does not eliminate variables. Moreover, it is unclear how well it might work in non-stationary processes.

There is a substantial literature on Bayesian model selection. Bayesian methods require the assumption of prior probabilities for the individual models or variables as well as prior distributions for the parameters, from which posterior probabilities can be

derived for the models and their parameters. Some Bayesian methods select a single model, for example, using the Schwarz (1978) information criterion (SIC), which is often referred to as the Bayesian information criterion as it was developed using priors, although Chow (1981a) notes that there is no need to justify SIC in terms of posterior probability criteria. Most Bayesian methods select a mixture of models by model averaging (see Hoeting, Madigan, Raftery and Volinsky, 1999, for an overview) in order to account for model uncertainty. Another aspect of Bayesian model selection is the extreme bounds analysis developed by Leamer (1978, 1983b, 1985) (also see Granger and Uhlig, 1990) who argues that inferences are only robust if the specification assumptions are broad enough to nest the DGP and the interval of inferences is narrow enough to be useful, which implies that the coefficient estimates should be statistically significant given some conventional decision rule. This requires prior information that Bayesians assume known. The approach has been heavily criticised by, *inter alia*, McAleer, Pagan and Volker (1985), Breusch (1990) and Hendry and Mizon (1990).

Simple-to-general modelling gradually augments the model until the preferred specification is achieved. Often, simplified theoretical models are postulated and subsequently tested against data. These models can fail for a number of reasons: there may not be a direct correspondence between the theoretical and empirical variables, time horizons may differ, latent variables may not be correctly measured, the theory may exclude relevant variables, exogeneity conditions may be violated, etc. If congruency tests are failed, the model is augmented until it passes the required tests. There are many problems with the technique, see Hendry and Krolzig (2001, p.132) for a critique. Primarily, the overall significance level of the strategy will not be controlled as there is no defined termination point, and the resulting model will be path dependent. The step-wise regression algorithm of Efroymson (1960) undertakes forward selection, in which variables are added to a set of selected variables and tests are implemented to see if any previously selected variable can be deleted without increasing the residual sum of squares. Conversely, the backward elimination procedure commences with the full set of regressors and eliminates one variable at a time (see Miller, 2002, ch.3, for an outline of Efroymson's algorithm). As

step-wise regression only searches one path, it does not have a high success rate of finding the DGP. Berk (1978) demonstrates that applying both forward and backward selection is no guarantee to finding the correct model, and Leamer (1983a) is highly critical of the step-wise approach. Alternative selection procedures exist, such as optimal regression in which all subsets of regressors are included (see Coen, Gomme and Kendall, 1969, and the response by Box and Newbold, 1971), but this approach is intractable with a large set of potential regressors. Numerous algorithms have been proposed to undertake an exhaustive search, see, *inter alia*, Schatzoff, Tsao and Fienberg (1968), Furnival (1971) and Morgan and Tatar (1972), but with little success.[2]

One of the most prominent approaches to model selection is that of general-to-specific modelling. The technique, which simplifies a general model that captures the salient characteristics of the data, has a long history and has been known as the LSE approach due to its proponents at the London School of Economics in the 1960s. Hendry (2003) discusses the origins of the LSE methodology and Mizon (1995) provides a history, while Pagan (1987) and Phillips (1988) provide reviews. Hansen (1996) critiques Hendry (1993), arguing that it is impossible to implement the *Gets* approach fully, and Faust and Whiteman (1995, 1997) refute the LSE approach due to the difficulty of deriving structural interpretations from reduced form models, to which Hendry (1997b) provides a response.

The idea behind the *Gets* methodology is that there is a complex mechanism that generates the observed data. The theory of reduction reduces this to a manageable framework, enabling empirical models to be derived from the DGP. The theory of reduction commences from the DGP given by:

$$\mathsf{D}_{\mathbf{u}}\left(\mathbf{U}_T^1|\mathbf{U}_0,\boldsymbol{\psi}_T^1\right), \quad \text{with} \quad \boldsymbol{\psi}_T^1 \in \boldsymbol{\Psi} \subseteq \mathbb{R}^{kT}, \tag{1.1}$$

where $\mathbf{U}_T^1 = (\mathbf{u}_1, ..., \mathbf{u}_T)$ is the full sample vector of random variables defined on probability space $(\boldsymbol{\Omega}, \mathcal{F}, \mathsf{P})$. As $\mathbf{U}_T^1$ is unmanageably large, the theory of reduction provides a series of data reductions to obtain the local data generating process (LDGP), which can

[2]Many procedures and algorithms have been proposed to narrow down the search space, including branch and bound techniques (see Hocking and Leslie, 1967, LaMotte and Hocking, 1970, and Gatu and Kontoghiorghes, 2006), ridge regression (see Hoerl and Kennard, 1970b, 1970a), the non-negative garrote (see Breiman, 1995), and the least absolute shrinkage and selection operator (LASSO: see Tibshirani, 1996).

subsequently be modelled. Initially, the parameters of interest are identified, $\boldsymbol{\mu} \in \mathcal{M}$, and the data are mapped to a new dataset, $\mathbf{U}_T^1 \leftrightarrow \mathbf{W}_T^1$, which includes any aggregation and transformations of the original data. The data are then sequentially factorized to obtain an innovation process:

$$\mathsf{D}_{\mathbf{w}}\left(\mathbf{W}_T^1|\mathbf{W}_0,\boldsymbol{\phi}_T^1\right)=\prod_{t=1}^{T}\mathsf{D}_{\mathbf{w}}\left(\mathbf{w}_t|\mathbf{W}_{t-1}^1,\boldsymbol{\delta}_t\right). \tag{1.2}$$

The data are partitioned into two sets, $\mathbf{X}_T^1$ and $\mathbf{V}_T^1$, where everything about $\boldsymbol{\mu}$ must be learnt from $\mathbf{X}_T^1$ alone (i.e. from $\boldsymbol{\delta}_{b,t}$).

$$\mathsf{D}_{\mathbf{w}}\left(\mathbf{w}_t|\mathbf{W}_{t-1}^1,\boldsymbol{\delta}_t\right)=\mathsf{D}_{\mathbf{v}|\mathbf{x}}\left(\mathbf{v}_t|\mathbf{x}_t,\mathbf{W}_{t-1}^1,\boldsymbol{\delta}_{a,t}\right)\times\mathsf{D}_{\mathbf{x}}\left(\mathbf{x}_t|\mathbf{X}_{t-1}^1,\mathbf{V}_{t-1}^1,\mathbf{W}_0,\boldsymbol{\delta}_{b,t}\right) \tag{1.3}$$

Marginalization then allows the conditional distribution to be discarded from (1.3), as well as the history of $\mathbf{V}_{t-1}^1$ in the marginal distribution if Granger non-causality is satisfied. The variables are then further partitioned into exogenous ($\mathbf{z}_t$) and endogenous ($\mathbf{y}_t$) variables via conditional factorization such that:

$$\mathsf{D}_{\mathbf{x}}\left(\mathbf{x}_t|\mathbf{X}_{t-1}^1\mathbf{W}_0,\boldsymbol{\delta}_{b,t}\right)=\mathsf{D}_{\mathbf{y}|\mathbf{z}}\left(\mathbf{y}_t|\mathbf{z}_t,\mathbf{X}_{t-1}^1,\mathbf{W}_0,\boldsymbol{\theta}_{a,t}\right)\times\mathsf{D}_{\mathbf{z}}\left(\mathbf{z}_t|\mathbf{X}_{t-1}^1,\mathbf{W}_0,\boldsymbol{\theta}_{b,t}\right), \tag{1.4}$$

in which the marginal distribution is discarded. Weak exogeneity ensures no loss of information in the reduction, which requires that the parameters of interest are a function of $\boldsymbol{\theta}_{a,t}$ alone and that the parameters are variation free, i.e. $(\boldsymbol{\theta}_{a,t},\boldsymbol{\theta}_{b,t}) \in \boldsymbol{\Theta}_a \times \boldsymbol{\Theta}_b$, see Engle, Hendry and Richard (1983). Further transformations, including mapping to stationarity, constancy, lag truncation and functional form, are undertaken to result in the derived LDGP, delivering the specification:

$$\mathbf{A}\left(L\right)\mathbf{g}\left(\mathbf{y}_t\right)=\mathbf{B}\left(L\right)\mathbf{h}\left(\mathbf{z}_t\right)+\boldsymbol{\varepsilon}_t \quad \boldsymbol{\varepsilon}_t \underset{app}{\sim} \mathsf{N}_n\left(\mathbf{0},\boldsymbol{\Sigma}_{\varepsilon}\right) \tag{1.5}$$

where $\boldsymbol{\varepsilon}_t$ is a mean-zero, homoskedastic, innovation process with variance $\boldsymbol{\Sigma}_{\varepsilon}$. Hendry (1995, ch.9) provides an outline of the theory of reduction.

Gets modelling attempts to mimic the theory of reduction in a practical setting. To summarise the approach, it is assumed that the DGP can be characterised by a sufficiently general model that nests the LDGP. A more parsimonious representation is

then searched for using a sufficiently exhaustive search procedure to ensure there is no path dependence. The final specific model should be statistically well-specified (denoted congruent), a valid restriction of the general model, and it should encompass every other model that is a valid restriction of the general regression.[3] Gilbert (1986, 1989) provides a detailed exposition of the *Gets* approach, Hendry (1993) outlines the progression of the methodology from its inception through a series of empirical papers, and Campos *et al.* (2005) provide an extensive discussion and review of the key papers in the literature. Also see Phillips (1988) and Ericsson, Campos and Tran (1990) for outlines of the approach.

Many criticisms have been levelled at the *Gets* methodology, ranging from the attack on data-based model selection and the work of Tinbergen (1940a, 1940b) by Keynes (1939, 1940) (see Hendry, 1980, and Hendry and Morgan, 1995, section 6) and the measurement without theory debate commenced by Koopmans (1947) (see Hendry and Morgan, 1995, p.69-71 and ch.43), to the more technical problems of pre-test biases, which arise because any statistical test has a non-zero size and non-unit power, but is a cost of inference as opposed to a cost of search; selection effects, whereby coefficient standard errors only reflect sampling variation conditional on a fixed model specification without taking into account model uncertainty; repeated testing, which delivers a high probability of obtaining spurious results due to undertaking many tests; a lack of identification, which requires a unique model that corresponds to reality; and path dependence of selection, which can be insured against by rigorous encompassing tests. Hendry (2000b, ch.20), Hendry and Krolzig (2001, ch.11), and Hendry and Krolzig (2003b) review and refute the range of criticisms. Furthermore, these criticisms can be addressed by rigorous analysis of the properties of a *Gets* model selection procedure, which is discussed in section 1.3 in relation to PcGets.

1.3 Automatic model selection

Given the range of model selection techniques discussed in section 1.2, it is unsurprising that a number of automatic algorithms have been developed to implement the techniques.

[3]The notion of encompassing has been extensively discussed in, *inter alia*, Mizon and Richard (1986), Hendry and Richard (1989), Mizon (1994) and Hendry (1995, ch.14).

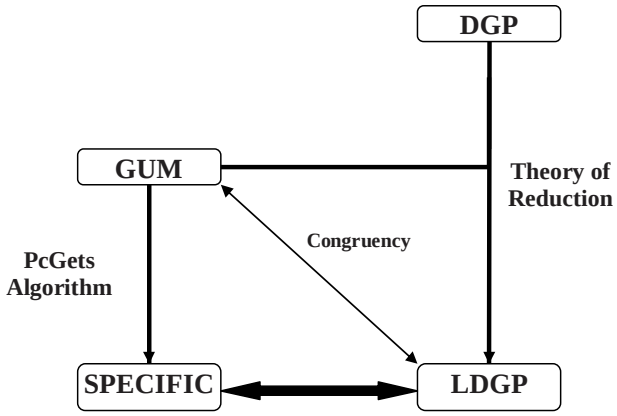

Figure 1.2: General-to-specific model selection

With regard to the *Gets* methodology, Hoover and Perez (1999) re-analysed Lovell (1983) using an algorithm that searched many different paths. The experimental design aimed to select a few (0-5) regressors from a large set of 40 variables. They reversed Lovell's conclusion that data mining was disastrous. We return to the Hoover and Perez experiment in Chapter 2 when we consider selection of nonlinear models. Hendry and Krolzig (1999) improve on Hoover and Perez's algorithm, developing the automatic model selection algorithm, PcGets. The econometric theory and methodology of the program are discussed in many publications including Hendry and Krolzig (1999, 2001, 2004a, 2004b, 2004d). Figure 1.2 conceptualises how the algorithm mimics the theory of reduction. Observe that the GUM must nest the LDGP defined in equation (1.5), which can be tested by congruency, in order for the algorithm to have a chance of locating the LDGP. The algorithm is outlined in detail in Chapter 2.

The properties of PcGets have been established by Krolzig and Hendry (2001) and Hendry and Krolzig (2003b, 2005), who demonstrate that PcGets retains relevant variables close to the theoretical maximum given by the power of a single `t`-test on the relevant variable, and eliminates irrelevant variables at the chosen significance level, such that with n irrelevant variables and a significance level of α, $n\alpha$ variables will be retained on average. Furthermore, extensive Monte Carlo evidence demonstrates that the equation standard error is close to that of the DGP, and while selected estimates are conditionally upward biased as the decision rule truncates the distribution to discard the mass near zero, coefficients have appropriate standard errors despite selection and can be

bias corrected using simple formulae, see Hendry and Krolzig (2005). Finally, Campos *et al.* (2003) show PcGets model selection to be consistent.

Various alternative automatic algorithms have been proposed in the literature; Hofmann, Gatu and Kontoghiorghes (2005) develop a computationally efficient automatic branch and bound algorithm, and substantial Bayesian model averaging software exists: see, e.g., the software developed by Adrian E. Raftery.[4] Other automatic algorithms implement selection based on information criteria, see, e.g., Phillips (1994, 1995). This algorithm is a data-based, automated model selection and forecasting tool that automatically detects non-stationarities and cointegrating relations and builds these into the model. The model selection criterion used is the posterior information criterion of Phillips and Ploberger (1994). Also see Phillips (1996, 2003) for comments. Geriner and Ord (1991) examine the use of automatic model selection techniques in developing multivariate forecasting models, and numerous algorithms exist to automatically select data-based models, including artificial neural networks, data mining algorithms, and dimension reduction methods.

1.4 Nonlinear models

Macroeconomic data comprises the aggregation and mapping of highly complex heterogeneous agents' actions. It is therefore almost certain that economic relationships are nonlinear in some form. There are numerous economic theories to support this claim, including regime-switching hypotheses, disequilibrium models, and production functions. However, most existing econometric models are linear, as it has proved difficult to develop tests for, and modelling of, general classes of nonlinear functions relevant in economics, given the practical sample sizes available in macroeconomics. Despite this precedence, there is an extensive literature looking at modelling nonlinearities (see, *inter alia*, Granger, 1993, Granger and Teräsvirta, 1993, and Teräsvirta, Tjostheim and Granger, 1994, for an overview), and many nonlinear model classes have met with success in explaining in-sample phenomena, although De Gooijer and Kumar (1992) argue that

[4] http://www.stat.washington.edu/raftery/software.html contains downloadable software. Also see Raftery, Painter and Volinsky (2005).

the evidence is ambiguous on whether the forecast performance of nonlinear models is superior to that of linear models, even if their in-sample fit is better.[5]

There have been a wide array of nonlinear models proposed in the literature, and our focus shall be on parametric models. Perhaps the most popular nonlinear models arise as extensions of linear autoregressive or moving average processes, see Granger and Teräsvirta (1993, ch.4) and references therein. Gallant (1987) provides an in-depth analysis of estimation and inference for nonlinear models, and Granger (1993) proposes a simple-to-general approach to selecting single variable nonlinear models. An extension of the nonlinear ARMA model is the bilinear model, which augments the ARMA model with the cross-product of the lagged error term and lagged dependent variable, see Priestley (1981, ch.11), Subba Rao and Gabr (1984) and Subba Rao (1985). As bilinear models are often difficult to invert (an essential requirement for forecasting) their use is limited, and they tend to be used for describing processes with occasional strong perturbances. Maravall (1983) shows that bilinear models perform no better than linear models in a forecasting context if there are no large perturbances in the forecast period.

A further class of models that captures the stochastic properties of economic data, but are in fact deterministic, are chaotic processes. Economic theory often postulates deterministic models, which is in direct contrast to evidence that data are stochastic, but Lorenz (1989) and Puu (1989) demonstrate that deterministic theories can lead to chaotic outcomes. Granger and Teräsvirta (1993, ch.3) find the evidence for economic data exhibiting chaotic behaviour is limited, and the techniques are not in common practice in econometrics.

There is a substantial literature on random coefficient and volatility models. Conditional heteroskedasticity models commenced with ARCH (see Engle, 1982, and the reviews in Engle and Bollerslev, 1986, Bollerslev, Chou and Kroner, 1992, and Bera and Higgins, 1993) and rapidly expanded to GARCH models (Bollerslev, 1986) and many others, see Engle (2002) for a review and look ahead to future developments in the field.

[5] On a note of caution, Granger and Teräsvirta (1999) provide an illustrative example of a case where criteria that have been developed for specifying linear models are of little or no use in a nonlinear context, and so linear properties are often not informative when describing the properties of a nonlinear process.

Tsay (1987) demonstrates that ARCH is a special case of random coefficient autoregressive models (RCA: see Nicholls and Quinn, 1982), which are, in turn, a subclass of time-varying parameter models (see Chow, 1984). Volatility models are now prominent in the literature, particularly in finance, with numerous applications; see Engle and Patton (2001) and Li, Ling and McAleer (2002) for surveys.

A wide class of models of particular interest are those of regime-switching models, including switching regression models of Goldfeld and Quandt (1973), Quandt (1983) and Hamilton (1989), as well as the class of smooth transition regression (STR) models (see Maddala, 1977, ch.7, Granger and Teräsvirta, 1993, ch.4 and ch.7, and Teräsvirta, 1994, 1998) which include threshold autoregression models (TAR: see Tong, 1978, 1990), smooth transition autoregression models (STAR: see Chan and Tong, 1986, Luukkonen, Saikkonen and Teräsvirta, 1988, Granger and Teräsvirta, 1993, and Teräsvirta, 1994), self-exciting threshold autoregressions (SETAR: see, e.g., Tong, 1983, and Clements, Franses, Smith and van Dijk, 2003), and exponential autoregression models (EAR: see Priestley, 1981, ch.11). These are popular tools for modelling state dependent behaviour. The STR model takes the form:

$$y_t = \boldsymbol{\beta}'\mathbf{X}_t + \left(\boldsymbol{\theta}'\mathbf{X}_t\right)\mathsf{G}\left(\gamma, c, s_t\right) + \epsilon_t, \quad \epsilon_t \sim \mathsf{IN}\left[0, \Omega\right], \tag{1.6}$$

for $t = 1, ..., T$, where s_t is the transition variable(s), c is the threshold, and γ controls the speed of transition. Various distributional assumptions can be made on $\mathsf{G}(.)$ providing it is bounded between 0 and 1: two popular choices are the logistic and exponential functions. A further generalisation is the STR-deviation model. The nonlinearity in a STR model is determined by the value of the transition function deviating from the fixed threshold c, whereas the nonlinearity in an STR-D model is generated by the deviations of a previous value or values of the dependent variable from a linear path. The evidence on forecasting with regime-switching models and smooth transition models is mixed, see Clements and Krolzig (1998) and Clements *et al.* (2003) for examples.

A final class of nonlinear models that have been popularised recently are artificial neural networks (ANN), which are a class of models that are designed to replicate the

way information is thought to be processed in the brain, see, e.g., White (1989) and Lee, White and Granger (1993). ANNs are a class of flexible nonlinear functions, or squashing functions, that are sufficiently general to approximate a wide range of nonlinear relationships. These models are primarily forecasting tools and are discussed in more detail in section 1.6.

Testing for functional form is an essential component of model building. There is an extensive literature on nonlinearity testing, including tests against a specific alternative and those against a more general alternative: Granger and Teräsvirta (1993, ch.6) provide an overview. Tests against a specific alternative can be formulated as Lagrange Multiplier (LM) tests, which means that estimation of the nonlinear model is not necessary, see Breusch and Pagan (1980). Furthermore, Pagan and Hall (1983) demonstrate that often LM tests are not necessary as it may be feasible to formulate nonlinearity tests as diagnostic tests on the residuals of the linear model.

Ramsey (1969) proposes a test for specification errors in regression, including unmodelled nonlinearity, based on adding powers of the fitted values: Doornik (1995) provides a careful evaluation of both the numerical and statistical properties of the RESET test. Keenan (1985) develops a univariate test for detecting nonlinearity (which is a special case of the RESET test) and Tsay (1986) extends Keenan's test to allow for contemporaneous nonlinearity. White (1980) develops a test that adds all squares of regressors, or squares and cross-products, to test for heteroskedasticity, implicitly testing for omitted nonlinearity as well. This has been investigated by numerous authors, including a recent appraisal in Hendry and Krolzig (2003b), relevant in our context of model selection.

Tests against specific alternatives include tests based on chaotic processes such as Brock, Dechert and Scheinkman (1987); tests against univariate bilinear models including Weiss (1986) and Saikkonen and Luukkonen (1988) (which is equivalent to testing against STR-D in the univariate case); and neural network tests proposed by Lee *et al.* (1993). Hinich (1982) and Ashley, Patterson and Hinich (1986) develop a bispectrum test that uses independence in the skewness to test for linearity. There are also a number of tests for nonlinearity in the variance, i.e. conditional heteroskedasticity, see Engle

(1982) and McLeod and Li (1983). Furthermore, Davidson and MacKinnon (1985) develop procedures to implement tests in regression directions (i.e., the null and, possibly implicit, alternative hypotheses are regression models of some sort and tests can therefore be computed as tests for omitted variables) in the presence of heteroskedasticity of an unknown form.

One of the principal difficulties in building nonlinear models in the time domain is to remain agnostic regarding the functional form, and yet to find a parsimonious parameter representation of sufficient generality to capture the underlying LDGP. Various approximating classes have been proposed, including, among others, polynomials, orthogonal polynomials, Fourier series, asymptotic series and confluent hypergeometric functions.

The use of polynomial functions has strong motivation, as a Taylor series expansion delivers a polynomial specification. A Taylor series expansion around 0 of an unknown functional form $y_t = f(\mathbf{W}_t, \mathbf{W}_{t-1}, ..., \mathbf{W}_{t-q}) + v_t$, where $\mathbf{W}_t$ is a distinct vector of n variables, will result in the dual of the Volterra series, see Priestley (1981, ch.11), given by:

$$\begin{aligned}\Psi(\mathbf{W}_t, ..., \mathbf{W}_{t-q}; \boldsymbol{\psi}) &= \psi_0 + \sum_{s=0}^{q}\sum_{i=1}^{n}\psi_{1,is}w_{i,t-s} + \sum_{r=0}^{q}\sum_{s=0}^{q}\sum_{i=1}^{n}\sum_{j=1}^{i}\psi_{2,ijsr}w_{i,t-s}w_{j,t-r} \\ &\quad + \sum_{p=0}^{q}\sum_{r=0}^{q}\sum_{s=0}^{q}\sum_{i=1}^{n}\sum_{j=1}^{i}\sum_{k=1}^{j}\psi_{3,ijksrp}w_{i,t-s}w_{j,t-r}w_{k,t-p} + ... \quad (1.7)\end{aligned}$$

This is a nonlinear generalisation of the Wold (1938) representation theorem, which states that the purely non-deterministic component of any stationary time-series can be represented as an infinite moving average of uncorrelated errors. The number of parameters expand rapidly, such that efficient estimation from a finite set of observations is impossible unless either some smoothness properties are assumed, severe truncation is applied, or the functions are expressed as known functions of a relatively small number of other parameters. Polynomials provide a good local approximation but their accuracy tends to decline further out. There are a substantial class of orthogonal polynomials of the form:

$$\mathsf{C}_n = \int_a^b w(x)\left[p_n(x)\right]^2 dx \quad (1.8)$$

where $p_n(x)$ is the class of polynomials defined over range $[a, b]$ and $w(x)$ is the weighting function. Various weighting functions can be imposed, including Hermite polynomials, with weighting e^{-x^2} over the domain $(-\infty, \infty)$; Chebyshev polynomials of the first kind, with weighting $(1 - x^2)^{-\frac{1}{2}}$ over the domain $[-1, 1]$; and Legendre polynomials, with weighting 1 over over the domain $[-1, 1]$: see, e.g., Abramowitz and Stegun (1972, ch.22) for details.[6] Rushton (1951) demonstrates how to orthogonalise using the Choleski method of solving linear equations. The concern with orthogonal polynomials as approximations is that they tend to perform poorly in the tails.

A Fourier series is an expansion of a periodic function in terms of an infinite sum of sines and cosines. The advantage of Fourier approximations is that they can capture the variation in any absolutely integrable function of time, such that the behaviour of any deterministic function can be readily captured by a sinusoidal function even though the function in question is not periodic, see Körner (1988). However, a close approximation requires many Fourier terms and the function also suffers from Gibbs phenomenon, see, e.g., Jerri (1998). Asymptotic expansions are discussed by Erdélyi (1987). In these functions, a series may converge or diverge but the partial sums can be an arbitrarily good approximation to the given function. Evidence on the ability of asymptotic series to approximate economic data is limited. Hypergeometric functions generalise many classes of functional forms, see Abadir (1999), and are becoming increasingly popular in economics. Hypergeometric functions are a general case of exponential functions, and confluent hypergeometric functions are a result of the product of exponential and polynomial functions. They can provide parsimonious general nonlinear estimation techniques when functional forms are unknown and they represent a wide range of classes of functions.

1.5 Inflation modelling

Inflation is a fundamental macroeconomic variable, driving both monetary and fiscal policy. A number of theories have been postulated to explain the determinants of inflation

[6] Other classes of orthogonal polynomials include, *inter alia*, Gegenbauer polynomials, Jacobi polynomials, Laguerre polynomials and generalised Laguerre polynomials.

(see Frisch, 1983, for a survey), with numerous empirical studies attempting to confirm or refute the various hypotheses. The development of contemporary inflation theory commenced with the Phillips (1958, 1962) curve, formalised by Lipsey (1960), which aimed to explain wage inflation by excess demand in the labour market. Empirical refutation led to the natural rate hypothesis, developed independently by Friedman (1968) and Phelps (1967), in which a distinction is made between the short-run and long-run Phillips curves, based on the 'natural rate' of unemployment. This was later challenged by Lucas (1973) and Sargent and Wallace (1975), who argued that rational expectations were more representative of agents behaviour than the adaptive expectations embodied in the natural rate model. Furthermore, Blanchard and Summers (1988) postulated the hysteresis hypothesis, which argues that the non-accelerating inflation rate of unemployment (NAIRU) is a function of actual employment. However, the expectations-augmented Phillips curve remains a dominant theory in mainstream macroeconomic analysis.

Monetarism reached its height in the early 1970s with Friedman's seminal works (Friedman, 1970, 1971), in which he argues that inflation is essentially a monetary phenomenon, i.e., if there is sustained money growth in excess of output growth, inflation will be generated (see Laidler and Parkin, 1975). To build inflation models of the quantity theory, the money-demand equation is inverted and nominal money is treated as exogenous, see Friedman (1956). Ericsson and Irons (1994) and Hendry (2000a) demonstrate the technical and empirical problems with this approach.

Excess demand and supply pressures in all sectors of the economy have provided explanations of inflation. In the goods market, final demands play a key role. The output gap, as a proxy for excess demand for goods and services, captures the difference between actual and potential, or trend, output. As a latent variable it is notoriously difficult to measure (see Hendry, 2000c) but numerous techniques have been proposed: for a useful survey see Dupasquier, Guay and St-Amant (1999). These range from linear and segmented linear trends; the Hodrick and Prescott (1997) filter (see Harvey and Jaeger, 1993, for a signal extraction interpretation and De-Brouwer, 1998, for a multivariate example); cubic splines (see, e.g., Green and Silverman, 1994, p.11: Doornik and Hendry,

1996, p.75, outline why the cubic spline and Hodrick Prescott filter result in similar decompositions); the Beveridge and Nelson (1981) filter (Proietti and Harvey, 2000, develop an algorithm for a two-sided filter and Evans and Reichlin, 1994, extend the filter to the multivariate case); band-pass filters (see, e.g., Baxter and King, 1999); structural time series models (see, e.g., Harvey and Jaeger, 1993: for a multivariate generalisation see Kuttner, 1994); structural VARS (see Blanchard and Quah, 1989); models based on the permanent income hypothesis (see Cochrane, 1994); to the production function method (for examples see, *inter alia*, Giorno, Richardson, Roseveare and van den Noord, 1995, Bolt and van Els, 2000, and Proietti, Musso and Westermann, 2002). However, the uncertainty associated with estimates of the output gap is substantial (see Orphanides and van Norden, 2002), particularly at the end of sample for univariate statistical techniques. De-Brouwer and Ericsson (1998), Hendry (2001), and Bowdler and Jansen (2004) all find the output gap to be an important determinant of inflation. Evidence for nonlinear inflation effects from the output gap have been examined in Turner (1995) and Clements and Sensier (2003), although the evidence for such effects is less robust.

Excess demand in the labour market or competition over the profit share captured by the mark-up often plays a prominant role in explaining inflationary pressure. Prevalent UK examples include, *inter alia*, Dicks-Mireaux and Dow (1959), Sargan (1964, 1980), Nickell (1990) and Layard, Nickell and Jackman (1991). The mark-up of prices over costs has a long history (see Duesenberry, 1950) and is often measured by unit labour costs and import prices. Nielsen and Bowdler (2003) find that unit labour costs feed through to the GDP deflator with a coefficient of 0.79, whereas Bårdsen, Fisher and Nymoen (1998) find a larger coefficient of 0.89, although this study excludes import prices and the real exchange rate. For further empirical analyses see, *inter alia*, Cockerell and Russell (1995), Bårdsen and Fisher (1999) and Batini, Jackson and Nickell (2000). With regard to labour demand pressures, the dominant contemporary macroeconomic theory is the NAIRU, see Ball and Mankiw (2002). Estimates of the NAIRU are as difficult to obtain as estimates of the output gap (see Staiger, Stock and Watson, 1997) but many empirical studies have estimated the NAIRU using a variety of approaches, see, *inter*

alia, Gordon (1997), Coen and Hickman (2002), and Batini and Greenslade (2006).

The approach to modelling inflation adopted in the thesis is embedded within the *Gets* framework, using equilibrium correction mechanisms (EqCM) to capture inflationary pressures from all sectors of the economy. EqCMs have become prevalent in modelling since their inception and development (see Granger, 1981, Engle and Granger, 1987, and Johansen, 1988), as most econometric models belong to the equilibrium correction class, including ADLs, VARs, DSGEs, and ARCH/GARCH models, see Hendry (2005a). Our focus will be on vector and single-equation EqCMs. For examples of equilibrium correction models of inflation in which many markets influence the inflation rate, see, *inter alia*, Rowlatt (1988), Juselius (1992), Metin (1995), De-Brouwer and Ericsson (1998), and Hendry (2001).

1.6 Forecasting

Economic forecasting has a long history, with serious criticisms thereof, stemming from the work of Morgenstern, reviewed in Hendry and Morgan (1995, p.17 and ch.13), (see Clements and Hendry, 1998b, ch.1.3, for a brief overview). Morgenstern argued that accurate forecasting was not feasible, both because economic data are not independently and identically distributed, and because forecasts are invalidated by agents' reactions to them. Marget (1929) provides an extensive discussion, refuting Morgenstern's conclusion. The first rigorous treatment of forecasting was provided by Haavelmo (1944), on which the framework for the theory of predictability is based, see Clements and Hendry (1996a, 1999, 2005a). Suppose there are T observations on a random variable, $\mathbf{X}_T^1 = (x_1, ..., x_T)$, from which to predict the future H values, $\mathbf{X}_{T+H}^{T+1} = (x_{T+1}, ..., x_{T+H})$. The joint probability is $\mathsf{D}_{\mathbf{X}_{T+H}^1}\left(\mathbf{X}_{T+H}^1 | \mathbf{X}_0, \boldsymbol{\theta}\right)$ where $\boldsymbol{\theta} \in \boldsymbol{\Theta} \subseteq \mathbb{R}^{\mathsf{p}}$ and $\mathbf{X}_0$ denotes the initial conditions. Factorizing into conditional and marginal probabilities results in:

$$\mathsf{D}_{\mathbf{X}_{T+H}^1}\left(\mathbf{X}_{T+H}^1 | \mathbf{X}_0, \boldsymbol{\theta}\right) = \mathsf{D}_{\mathbf{X}_{T+H}^{T+1}}\left(\mathbf{X}_{T+H}^{T+1} | \mathbf{X}_T^1, \mathbf{X}_0, \boldsymbol{\theta}\right) \times \mathsf{D}_{\mathbf{X}_T^1}\left(\mathbf{X}_T^1 | \mathbf{X}_0, \boldsymbol{\theta}\right). \tag{1.9}$$

As $\mathsf{D}_{\mathbf{X}_{T+H}^{T+1}}(.)$ is unknown at T, it must be estimated from $\mathsf{D}_{\mathbf{X}_T^1}(.)$, which requires the specification of $\mathsf{D}_{\mathbf{X}_T^1}(.)$ to imply the complete specification of $\mathsf{D}_{\mathbf{X}_{T+H}^1}(.)$. In this case, the only uncertainty in the forecast will be due to predictable uncertainty of the innovation

process generated by the sequential factorization and the estimation uncertainty from $\widehat{\boldsymbol{\theta}}$, see Ericsson (2002). However, in practice the mapping from $\mathsf{D}_{\mathbf{X}_1^T}(.)$ to $\mathsf{D}_{\mathbf{X}_{T+H}^{T+1}}(.)$ is not direct. Model mis-specification, data mis-measurement and structural breaks all impact on the ability to forecast future outcomes, and substantial empirical evidence of the forecast failure of econometric models in forecasting competitions is provided in, e.g., Makridakis and Hibon (2000), Clements and Hendry (2001), and Fildes and Ord (2002), all of which point to these factors as the fundamental components of explanations of forecast failure.

The theory of predictability provides the necessary, but not sufficient, conditions for forecastability, see Clements and Hendry (2005a). A random variable ν_t is unpredictable with respect to an information set $\mathcal{I}_{t-1}$ over a period $\mathcal{T}$ if its conditional distribution $\mathsf{D}_{\nu_t}(\nu_t|\mathcal{I}_{t-1})$ equals its unconditional $\mathsf{D}_{\nu_t}(\nu_t)$:

$$\mathsf{D}_{\nu_t}(\nu_t|\mathcal{I}_{t-1}) = \mathsf{D}_{\nu_t}(\nu_t), \quad \forall t \in \mathcal{T}. \tag{1.10}$$

Thus, a predictable process requires some combination with the information set:

$$\mathbf{y}_t = \phi_t(\mathcal{I}_{t-1}, \nu_t), \tag{1.11}$$

such that $\mathsf{D}_{\mathbf{y}_t}(\mathbf{y}_t|\mathcal{I}_{t-1}) \neq \mathsf{D}_{\mathbf{y}_t}(\mathbf{y}_t)$, $\forall t \in \mathcal{T}$. Predictability is relative to the information set used, for example, if $\mathcal{J}_t \subset \mathcal{I}_t$ the case could occur where $\mathsf{D}_{\nu_t}(\nu_t|\mathcal{J}_{t-1}) = \mathsf{D}_{\nu_t}(\nu_t)$ and yet $\mathsf{D}_{\nu_t}(\nu_t|\mathcal{I}_{t-1}) \neq \mathsf{D}_{\nu_t}(\nu_t)$. However, while the predictions from a proper subset of information (i.e. contains the $\sigma-$field generated by $(\mathbf{y}_{t-1}, ...)$) will be less accurate, they will remain unbiased. Increasing the information set by disaggregation is addressed in Chapter 5 (see Espasa, Senra and Albacete, 2002, and Hubrich, 2005, for empirical studies). A further important result from the theory of predictability highlights the problems that structural breaks can cause. As unpredictability is relative to the historical time period, it is possible that $\mathsf{D}_{\nu_t}(\nu_t|\mathcal{I}_{t-1}) \neq \mathsf{D}_{\nu_t}(\nu_t)$ for $t = 1, ..., T$, but $\mathsf{D}_{\nu_t}(\nu_t|\mathcal{I}_{t-1}) = \mathsf{D}_{\nu_t}(\nu_t)$ for $t = T+1, ..., T+H$. Thus, there is no guarantee that a model that explains the phenomena of interest in-sample will be able to forecast out-of-sample. Clements and Hendry (1999) provide a taxonomy of forecast errors to determine all potential sources

of forecast error, including both the requirements for predictability, and the subsequent requirements for forecastability.

There are numerous techniques and methods for economic forecasting, ranging from simple rules of thumb and extrapolation to large econometric systems. Texts that provide an overview include Granger and Newbold (1986), Harvey (1989), Engle and White (1999), and Hendry and Ericsson (2001). Surveys of consumers' and businesses' plans and expectations provide useful information regarding future events, but require the plans to be realised and are uninformative if alternative plans are realised, see Nerlove (1983). However, surveys do play a role in augmenting the information set and can be included in econometric models to capture latent variables.[7] Granger (1989, ch.6) provides an overview of the use of survey data.

Leading indicators are another conventional forecasting tool, popularised by the NBER over 50 years ago. A composite leading index of economic activity aims to predict economic expansions and contractions, but many constancy assumptions are required regarding both the variables in the index and the weights on the variables, for the indicator to be of use. In practice, the composition of leading indicators is changed regularly, see, *inter alia*, Diebold and Rudebusch (1989, 1991), Stock and Watson (1989, 1993), and Camba-Mendez, Kapetanios, Weale and Smith (2002) for examples, and Emerson and Hendry (1996) and Clements and Hendry (1998b, ch.9) for an evaluation. Lahiri and Moore (1991) and the papers therein provide an extensive overview.

Structural time series models and Box-Jenkins approaches are closely interlinked as there is a direct mapping between unobserved components models and ARIMA models, see Harvey (1989). The Box and Jenkins (1970) framework has enjoyed relative popularity since its inception, primarily due to the Wold (1938) decomposition theorem, but also due to its apparent empirical success relative to econometric models. Subsequently, the ARMA model has provided the standard benchmark against which forecasts are measured: see, *inter alia*, Naylor, Seaks and Wichern (1972), Granger and Newbold (1986,

[7]For example, the Bank of England makes use of surveys such as the Chartered Institute of Purchasing and Supply survey for services data, the British Chambers of Commerce surveys for capacity utilisation, and the NOP inflation attitudes survey, among others (see, e.g., Britton, Cutler and Wardlow, 1999, and Harrison, Nikolov, Quinn, Ramsay, Scott and Thomas, 2005).

ch.5) and Pedregal and Young (2002). The success of these simple parsimonious models is partly due to their relative robustness to structural breaks and partly due to mis-specification of the comparative econometric models. Unobserved components models are the modern successor of Box-Jenkins models (see, *inter alia*, Harvey and Todd, 1983, Andrews, 1994, and Proietti, 2002). The approach has wide appeal as the Kalman (1960) filter enables time-varying parameter models to be estimated, interpreted and forecast directly using a program such as STAMP (Koopman *et al.*, 1995), although the forecasts will not be robust if the time-varying trend is poorly estimated or subject to structural breaks.

Granger and Newbold (1986), Hamilton (1994, ch.4) and Diebold (1998) all provide overviews of univariate time-series forecasting approaches. The ARIMA model has been generalised in many directions. In the univariate context, there is a substantial literature on volatility forecasting, with a focus on conditional variance models such as ARCH (Engle, 1982), GARCH (Bollerslev, 1986), IGARCH (e.g. Lamoureux and Lastrapes, 1990), EGARCH (Nelson, 1991), and nonlinear GARCH (e.g. Glosten, Jagannathan and Runkle, 1993, and Lanne and Saikkonen, 2005). Long-memory models such as ARFIMA (Baillie, Bollerslev and Mikkelsen, 1996) have also been used to forecast. Anderson and Bollerslev (1998) and Christoffersen and Diebold (2000) assess whether volatility is forecastable, and numerous studies have had varying degrees of success.

Neural networks are relatively new in the field of economics, introduced by Kuan and White (1994): Kohzadi, Boyd, Kaastra, Kermanshahi and Scuse (1995) provide a useful introduction. As they allow for very general nonlinear relationships between variables they are a popular forecasting tool, but success has been limited, see Swanson and White (1995, 1997b) and Angstenberger (1996) in macroeconomics and finance respectively.

Evidence on the forecasting performance of nonlinear models has been mixed at best: see, *inter alia*, Granger and Teräsvirta (1993, ch.8), Brooks (1997), Stock and Watson (1999b), Clements, Franses and Swanson (2004) and Granger (2005). Furthermore, obtaining optimal forecasts for more than 1-step ahead is very difficult for nonlinear models, see Granger and Lin (1994) who consider multi-step forecasts from a nonlinear

autoregressive time-series model. However, many studies have delivered favourable results. For example, Dahl and Hylleberg (2004) examine the forecast performance of four nonlinear models and find evidence that some of the flexible nonlinear regression models perform well relative to the nonlinear benchmark, and Marcellino (2004) finds that nonlinear models often perform better than linear models when forecasting aggregate EMU macroeconomic variables in a real-time forecasting framework. Furthermore, Clements and Galvão (2004) and Sensier, Artis, Osborn and Birchenhall (2004) demonstrate the relevance of nonlinear models for forecasting.

The multivariate extensions to Box-Jenkins models are the VAR and VARMA (see Quenouille, 1957, and Tiao, 1981, with a review in Lütkepohl, 1991). Engle and Yoo (1991) demonstrate that a VARMA representation can be derived from the Wold (1938) decomposition theorem. Sims (1980) is a strong advocate of the VAR approach, which has become popular for forecasting (see, e.g., Doan, Litterman and Sims, 1984). However, Clements and Hendry (1999) demonstrate that the detectability of any structural change, particularly in the deterministic terms, is not easily reflected by the VAR parameterisation. Developments in cointegration analysis have led to EqCMs as the dominant class of econometric forecasting models. Engle and Yoo (1987) and Clements and Hendry (1995) demonstrate that VEqCMs should outperform VARs in a forecasting context, but the class of models is not robust to structural breaks either, which is highlighted by the forecast failure of many empirical equilibrium correction models; see, e.g., Hendry and Mizon (1993) for UK money demand and Clements and Hendry (1998a) for UK consumers' expenditure.

Forecasting using large scale macro-econometric models requires different techniques and evaluation criteria to standard models due to the high dimensionality and nonlinearity of such systems, see Clements and Hendry (1998b, ch.7). For applications, see Fair (1984) for the US, and Hatch (2003) who discusses the Bank of England's core model. Korenok and Swanson (2005) examine the forecast performance of DSGE models. An alternative approach that is becoming more popular is factor forecasting. The approach stems from principal components analysis developed by Hotelling (1933) and outlined

in Muirhead (1982) and Anderson (1994), and is implemented in the form of diffusion indices (see, e.g., Stock and Watson, 1998, 2002) and dynamic factor models (see, e.g., Forni, Hallin, Lippi and Reichlin, 2000). Bernanke and Boivin (2003) argue that factor models have the advantage of offering a framework for analysing data that is clearly specified and statistically rigorous, but that remains agnostic regarding the structure of the economy. Hendry and Clements (2005) examine the properties of factor forecasts in the presence of structural breaks: if the factors pool the information set, a smaller error variance could be obtained, but if the weights on the factors change in ways that are not captured, the factor forecasts perform poorly. Stock and Watson (1999c) provide an empirical example, demonstrating that principal component forecasts outperform a standard Phillips curve model when forecasting US inflation at the 12−month horizon.

Clements and Hendry (1999, ch.3) demonstrate that structural breaks are the most pernicious cause of forecast failure, particularly shifts in the deterministic terms (also see Clements and Hendry, 2002b, 2003). Robust forecasting methods have been developed to overcome the forecast failure associated with these breaks, see, *inter alia*, Hendry and Clements (2000) and Hendry (2005b). Various techniques have been proposed, including differencing, intercept corrections, rapid updating and pooling. Differencing ensures that the impact of shifts in deterministic terms are reduced by lowering the degree of the polynomial in time, eliminating shifts in both trend and location. Figure 1.3, taken from Hendry (2005b), demonstrates this: double-differencing deterministic terms reduces a location shift to a blip and a trend break to an impulse. A double differenced device is based on the premise that most economic time-series do not accelerate continuously, and hence $\mathsf{E}\left[\Delta^2 \mathbf{x}_t\right] = \mathbf{0}$, so the forecasting rule given by:

$$\widetilde{\Delta \mathbf{x}}_{T+1|T} = \Delta \mathbf{x}_T \tag{1.12}$$

delivers unconditionally unbiased forecasts. The rule captures the LDGP even if the econometric model is mis-specified, albeit with a 1-period lag, and this helps to explain the device's success. A second robust device, proposed by Clements and Hendry (2005a) and Hendry (2005b), differences the VEqCM, thereby removing the deterministic terms,

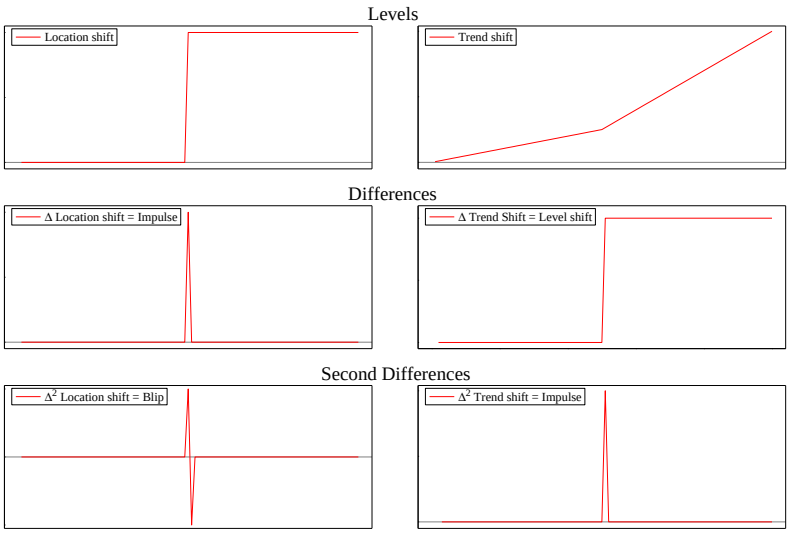

Figure 1.3: Location shifts and broken trends

see Chapter 5 for details. This device robustifies the model to structural breaks while still retaining the causal information through the cointegrating relations, although there is a cost incurred due to an increased error variance.

Model-based forecasts, in which no adjustments are applied, are rarely reported. Judgements or corrections are often used, either to set the model 'back on track', or to capture anticipated future events that are not explicitly modelled, see Clements and Hendry (1998b, ch.8). Forecast adjustments are common: see, *inter alia*, Klein (1971, p.48), Klein, Howrey and MacCarthy (1974), Wallis, Fisher, Longbottom, Turner and Whitley (1987, ch.4.3), and Turner (1990), and they have appeared to improve forecasting performance, see, e.g., Wallis and Whitley (1991). However, Clements and Hendry (2005b) demonstrate that intercept corrections can worsen forecast performance, depending on the underlying LDGP, and hence caution should be applied when using these techniques.

Exponential smoothing has a long history stemming from Holt (1957) and Winters (1960). The exponentially weighted moving average (EWMA: see Gardner, 1985) is a smoothing device, given by the recursive updating formula:

$$\widehat{y}_{T+1|T} = (1-\lambda)\sum_{j=0}^{\infty}\lambda^{j}y_{T-j}, \tag{1.13}$$

for $\lambda \in (0,1)$, such that:

$$\widehat{y}_{T+1|T} = (1-\lambda)\, y_T + \lambda \widehat{y}_{T|T-1} = y_T - \lambda \left(y_T - \widehat{y}_{T|T-1}\right) \tag{1.14}$$

with $\widehat{y}_1 = y_1$. The EWMA can be seen as approximating an ARIMA(0,1,1) and is therefore often used to overcome measurement error. Clements and Hendry (2005a) observe that the apparent success of the technique is due to a number of factors including (i) the adaptive behaviour of updating the next forecast by the previous forecast error; (ii) differencing to avoid location shifts; (iii) the lack of deterministic terms; and (iv) the rapid adaption when λ is small. However, the EWMA dampening factor has the opposite impact to intercept corrections, and so there is a dichotomy between accounting for breaks and allowing for measurement error.

Forecast combination or pooling is extremely common in practical forecast applications and often produces superior forecasts to the individual forecasts: Clemen (1989) provides a review, and surveys are given by Diebold and Lopez (1996) and Newbold and Harvey (2002). Bates and Granger (1969) propose the combination of two or more forecasts using:

$$\widehat{y}_{T+h} = \alpha \widehat{y}_{1,T+h} + (1-\alpha)\, \widehat{y}_{2,T+h}, \tag{1.15}$$

where $\widehat{y}_{i,T+h}$ is the h-step ahead forecast made at T using model i, for $i = 1, 2$. The in-sample period is used to calculate the optimal weights and various approaches to estimating $\widehat{\alpha}$ have been proposed. Bates and Granger (1969) use a variance-covariance approach, whereas Granger and Ramanathan (1984) propose obtaining α from the regression of forecast errors on the difference between the two models' forecasts, see Diebold (1988). Hendry and Clements (2004) propose various explanations for the success of forecast pooling. If the individual forecasts are based on different subsets of the full information set, or if forecasts are differentially biased, a combination may improve the forecast accuracy. Forecast pooling may also offset the impact of structural breaks and it may reduce the variance if different information sets are used. Furthermore, combination may act like an intercept correction. Finally, pooling can also be viewed as a shrinkage estimation technique, which may have a closer correspondence to the underlying reality.

Multi-step forecasts are essential to economic forecasting, and yet there are relatively few formal analyses of the properties of multi-step forecasts; see Clements and Hendry (1996b) and Bhansali (2002). To forecast h-steps ahead, either an iterated 1-step estimator or a direct h-step estimator can be used. Consider an h-step ahead forecast from a VAR for the n variables $\mathbf{y}_t$:

$$\mathbf{y}_t = \mathbf{\Pi}\mathbf{y}_{t-1} + \boldsymbol{\epsilon}_t, \tag{1.16}$$

where $\mathsf{E}[\boldsymbol{\epsilon}_t] = \mathbf{0}$. The iterated 1-step forecast is defined as:

$$\widehat{\mathbf{y}}_{T+h} = \widehat{\mathbf{\Pi}}^h \mathbf{y}_T, \tag{1.17}$$

where the forecast error is $\widehat{\boldsymbol{\epsilon}}_{T+h|T} = \left(\mathbf{\Pi}^h - \mathsf{E}\left[\widehat{\mathbf{\Pi}}^h\right]\right)\mathbf{y}_T$. The direct h-step estimator is:

$$\widetilde{\mathbf{y}}_{T+h} = \widetilde{\mathbf{\Pi}}_h \mathbf{y}_T, \tag{1.18}$$

where
$$\widetilde{\mathbf{\Pi}}_h = \arg\min_{\mathbf{\Pi}_h} \left| \sum_{t=1}^{T} \left(\mathbf{y}_t - \mathbf{\Pi}_h \mathbf{y}_{t-h}\right)\left(\mathbf{y}_t - \mathbf{\Pi}_h \mathbf{y}_{t-h}\right)' \right|, \tag{1.19}$$

such that the forecast error is $\widetilde{\boldsymbol{\epsilon}}_{T+h|T} = \left(\mathbf{\Pi}^h - \mathsf{E}\left[\widetilde{\mathbf{\Pi}}_h\right]\right)\mathbf{y}_T$. The relative forecast accuracy of the two forecasts depends on the estimators, $\widehat{\mathbf{\Pi}}^h$ and $\widetilde{\mathbf{\Pi}}_h$. Chevillon and Hendry (2005) find that the iterated 1-step forecasts are preferable if the model is well-specified, for both stationary and I(1) processes, but if there is mis-specification, deterministic shocks or negative serial correlation, direct multi-step estimation may lead to more accurate forecasts. Clements and Hendry (1998b, ch.11) confirm these results, noting that model mis-specification is necessary but not sufficient to justify direct estimation. The conclusions suggest limited use of direct estimation techniques, although the empirical results of Tsay (1993) and Lin and Tsay (1996) are more positive regarding direct h-step forecasting.

Clements (2005, p.1) defines a forecast to be any statement about the future, which implies the type of forecast can range from a point forecast to a density forecast, over any future horizon. Point forecasts have traditionally dominated the literature, sparking a vast literature on the evaluation of point forecasts, see Granger and Newbold (1973), Clements and Hendry (1998b, ch.3), Granger (2003) and Clements (2005) for critiques. The ideal evaluation criterion would depend on the cost function of the user of the

economic forecast, and so denoting the forecast error as $e_{T+h} = y_{T+h} - \widehat{y}_{T+h}$, where the forecast is $\widehat{y}_{T+h}$, the point forecast should solve:

$$\min_{y_{T+h}} \int_{-\infty}^{\infty} C\left(\widehat{y}_{T+h} - x\right) p_{T+h}\left(x|\mathcal{I}_{T+h}\right) dx \tag{1.20}$$

where the cost function is $C(.)$, and $p_{T+h}\left(x|\mathcal{I}_{T+h}\right)$ is the conditional probability density function. Granger and Pesaran (2000a, 2000b), Pesaran and Skouras (2002) and Clements (2005, ch.6) discuss forecast evaluation in the framework of decision-based methods. In practice, specifying the user's cost function is difficult and a squared cost function is often postulated, $C\left(e\right) = e^2$, which implies the optimal point forecast is the conditional mean. This results in the MSFE criterion. While popular, the MSFE criterion may not result in a definitive ranking as the measure is not invariant to non-singular, scale-preserving linear transformations, see Clements and Hendry (1993a, 1993b). Hence, different forecast rankings can be obtained if different isomorphic representations of a process are examined. Furthermore, rankings can differ over h-step forecasts, where $h > 1$, as the MSFE criterion does not account for the covariances between the forecasts at different horizons. Clements and Hendry (*op. cit.*) propose the generalised forecast-error second-moment criterion (GFESM), based on the determinant of the forecast-error second-moment matrix, $\mathbf{\Phi}_h = \mathsf{E}\left[\mathbf{E}\mathbf{E}'\right]$, obtained by stacking all previous step ahead forecast errors, $\mathbf{E}' = \left[e'_{T+1}, e'_{T+2}, ..., e'_{T+h}\right]$.

Tests of forecast accuracy are used for both testing predictive failure and for comparing rival forecasts. A test of predictive accuracy is equivalent to testing for structural change in the form of non-constant parameter estimates and error variances across two sub-samples, see Chow (1960), Christ (1966, ch.10), Hendry (1979) and Hackl and Westlund (1991): Pesaran, Smith and Yeo (1985) and Mariano (2002) provide reviews. Box and Tiao (1976) formalise a test that compares the estimate of the forecast error variance obtained from past residuals with the actual MSFE. Tests of comparative forecast accuracy attempt to assess whether rival forecasts are significantly different or whether the differences can be attributed to sampling variability, see Hendry (1986) and Andrews, Minford and Riley (1996). Forecast encompassing tests, which test whether one model

can explain the forecast errors made by an alternative model (where Chong and Hendry, 1986, present the test as a model specification test for large-scale models), are formally equivalent to testing for 'conditional efficiency' (see Nelson, 1972, and Granger and Newbold, 1973), where the forecast is conditionally efficient if the variance of the forecast error from a combination of two forecasts is not significantly less than that of the original forecast alone. Details of encompassing tests can be found in Harvey, Leybourne and Newbold (1998, 2000).[8] The established test statistic to evaluate the equality of MSFEs is derived by Granger and Newbold (1986) and is often referred to as the Morgan-Granger-Newbold test due to Morgan (1940). Diebold and Mariano (1995) discuss relaxing the assumptions of zero-mean, normally distributed and serially uncorrelated errors, and Harvey, Leybourne and Newbold (1997) refine the test to relax normality.

The forecast of a random variable should be described in terms of its conditional distribution, which provides a complete description of the uncertainty associated with the forecast. Point forecasts limit the information by looking at the mean or median estimate, excluding any measure of uncertainty. Density forecasts were relatively uncommon until the last decade, see Tay and Wallis (2002) for a survey, but they have become popular both in finance (see, *inter alia*, Diebold, Hahn and Tay, 1999, Granger and Sin, 2000, and Berkowitz, 2001, for both estimation and evaluation) and in macroeconomics, most explicitly observed in the "rivers of blood" and "rivers of bile" fan charts produced by the Bank of England (see Coyle, 2001).[9] Interval forecasts collapse the information in a density forecast to give a range for the forecast, with an assigned probability (see Chatfield, 1993, and Clements, 2005, ch.4) and are particularly popular in Value-at-Risk analysis. Lopez (1999) provides a discussion and Wallis (2005b) draws together density and interval forecasting.

[8]The impact of parameter estimation uncertainty on tests of forecast encompassing, and other tests of predictive accuracy when the forecasts are model-based, are examined by West (1996, 2001) and West and McCracken (1998).

[9]The "rivers of blood" refer to the Bank of England inflation forecasts, produced as a fan chart in shades of red, and the "rivers of bile" refer to GDP forecasts, produced in green. For more details on the Bank of England forecasts of inflation and output see Britton, Fisher and Whitely (1998) and Wallis (1999, 2005a). An evaluation using the Kullback-Leibler Information Criterion is given in Mitchell and Hall (2005).

1.7 Conclusion

The fundamental concern in model selection is to assess the costs of retaining irrelevant variables and the costs of excluding relevant variables. These costs can be calibrated for in-sample models (e.g., Hendry and Krolzig, 2005, calculate the costs for PcGets) but less is known in a forecasting context. Hendry and Hubrich (2006) calculate the MSFE, both including and excluding a set of relevant regressors for a static linear regression model, and express these in terms of the non-centralities of the excluded variables. They show that variables with a squared non-centrality greater than one should be retained, which implies a much looser selection criterion than is used to select in-sample models (it is, however, equivalent to the implicit significance level of AIC at $T = 100$ for a range of n). This is in contrast to the nonlinear model selection literature that suggests strong evidence of nonlinearity is required before building nonlinear forecasting models.

The literature regarding the empirical modelling and forecasting of inflation is abundant, and the recent progress made in the fields of automatic model selection, understanding of forecast failure, and the development of nonlinear models has led to a broad knowledge base on which this thesis builds, in keeping with a progressive research strategy.

PART I

MODEL SPECIFICATION

Chapter 2

Evaluating PcGets and RETINA as Automatic Model Selection Algorithms

2.1 Introduction

Advances in automatic model selection procedures have been swift in recent years, with impressive results. Despite the controversy surrounding many model selection strategies, and the multitude of diverse approaches in the literature, automatic procedures have been developed that show remarkably 'good' properties. Two such procedures are PcGets and RETINA (relevant transformation of the inputs network approach). PcGets is based on a general-to-specific search strategy, starting with a general model capturing the underlying characteristics of the data and testing downwards, ensuring validity of the reductions at each stage to result in a congruent, parsimonious, undominated model. RETINA differs from PcGets in that the *Gets* methodology is not its main tenet. The program uses a specific-to-general approach whereby variables are added into the model depending on a given criterion. RETINA aims to identify a parsimonious set of variables that are likely to be relevant for predicting out-of-sample.

This chapter brings together the two selection algorithms, reviewing their differing structures and objectives. RETINA automatically generates nonlinear transformations and then uses disjoint sub-samples to select, performing a selective search informed by out-of-sample fit criteria. PcGets requires the user to formulate a general model, performing an exhaustive search over the full sample using diagnostic and significance testing as the selection criteria. An assessment of these algorithms on both time-series and cross-section data is informative as to the comparative performance of the two programs. Furthermore, extensive Monte Carlo analysis exposes the properties of the two selection algorithms when selecting nonlinear models. This analysis is requisite in establishing the use of automatic algorithms to select models for the purpose of forecasting, and will help

to inform the econometrician as to the most appropriate approach to use when selecting models, depending on the underlying characteristics of the data.

The chapter is structured as follows. Section 2.2 outlines the RETINA algorithm and discusses its objectives and performance claims. Section 2.3 outlines the PcGets algorithm and notes the development that PcGets can handle more variables than observations. Section 2.4 reviews the differences between the two approaches. Section 2.5 examines the performance of the two programs for applications including the cross-section data analysed in Perez-Amaral, Gallo and White (2005), and the time-series data analysed in Hoover and Perez (1999). Section 2.6 presents a range of Monte Carlo experiments assessing the null and non-null rejection frequencies of both of these automatic model selection procedures when searching over nonlinear functions, for orthogonal and non-orthogonal models. Finally, section 2.7 concludes, reviewing the importance of automatic model selection procedures and assessing their use in a nonlinear framework.

2.2 RETINA

RETINA is a method of model selection along the lines of neural network models and is designed to identify a parsimonious set of regressors to predict out-of-sample. The algorithm is outlined in Perez-Amaral, Gallo and White (2003). RETINA is useful when the functional form of the conditional mean of the dependent variable is unknown, testing for nonlinearities and interaction effects within the modelling procedure. Concavity of the likelihood is achieved by imposing linearity in the parameters. The method relies on a sub-sample cross-validation scheme to ensure parsimony. Rather than undertaking an exhaustive model search, which would require the evaluation of 2^m models for m variables, the number of models is narrowed down by including variables sequentially in rank order. Collinearity is controlled by ensuring the R^2 between the included variables and the additional variable lies below a specified threshold parameter, λ. This section describes RETINA and briefly discusses how it works.

2.2.1 The RETINA selection algorithm

There are four main stages in the RETINA algorithm, including data building and sorting, isolating a candidate model, the search strategy and model selection. Initially, a set of potentially relevant variables, $\mathbf{X}$, that are thought to contain information about the conditional mean of the dependent variable, y, are identified and labelled 'level 0 transforms'. A set of transformed variables, $\zeta(\mathbf{X}) = \{W_1, ..., W_M\}$, are generated and denoted 'level 1 transforms'. There are numerous possible transformations capturing both nonlinearities and interactions, and RETINA uses the transformations given by:

$$\zeta(\mathbf{X}) = X_{i,h}^{\alpha} X_{i,j}^{\beta}, \quad \text{for} \quad \alpha, \beta = -1, 0, 1, \tag{2.1}$$

which results in a set of M potential predictors. Note that further iterations of the transformations can be implemented and appended to the level 0 and level 1 transforms.[1]

The sample is randomly divided into three disjoint sub-samples, N_k for $k = 1, 2, 3$, from which a candidate model is selected. This is, of course, appropriate only for cross-section data. As there are six pairwise permutations of the three sub-samples, a candidate model is chosen for each ordering. To isolate a candidate model for one of the permutations, the variables in the first sub-sample, W_{kj} for $j = 1, ..., M$, are ranked according to a relevance measure. RETINA uses the sample correlation with y, denoted $|\widehat{\rho}_{kj}|$. Starting with an intercept and the first variable, W_{k1}, which has the highest absolute sample correlation, successive variables are added in their ranked order, ensuring the R^2 of the regression of the last added variable on the subset already included lies below a specified threshold parameter, λ $(0 < \lambda < 1)$, i.e. include W_{kj} if $\mathsf{R}^2_{kj} \leq \lambda_p$ where R^2_{kj} is derived from the regression of:

$$W_{kj,i} = \beta_0 + \sum_{l=1}^{j-1} \beta_l W_{kl,i} + \epsilon_i \quad \text{for} \quad i = 1, ..., N_k. \tag{2.2}$$

This results in a set of variables selected on the basis of λ_p. Repeating the process for a grid of values for λ, a set of models will be retained, denoted by $\zeta_p(\mathbf{X})$ for $p = 1, ..., \nu$,

[1]The transformations include squares, inverses, squared-inverses, cross-products, cross-inverses and ratios. Some transformations may be excluded in practice because of numerical problems. For example, the inverse of X_j will be excluded if at least one observation is equal to 0.

where p determines the value of λ.[2]

MSFE, given by $N_k^{-1} \sum_{i=1_k}^{N_k} (y_i - \widehat{y}_i)^2$ where k denotes sub-sample k, is used to select between the ν potential models. The models are estimated in the first sub-sample and the MSFE is computed for the second sub-sample. The candidate model is chosen on the basis of lowest MSFE in the second sub-sample and is denoted the 'local best model'. This model will have a corresponding optimal λ^*. MSFE criterion is used as a means of sub-sample cross-validation to focus the selection algorithm on out-of-sample predictive ability.

The third stage of the algorithm aims to select a more parsimonious model by searching over all other possible models. Using the regressors in the local best model, models are estimated in the second sub-sample by adding in regressors sequentially, starting with the highest ranked variable. This results in S models, where S is the number of regressors in the local best model. These regressors are also ranked based on correlations with the dependent variable in the second sub-sample and are again included sequentially, giving $S-1$ models if the rankings differ. Out-of-sample AIC is computed for all $2S-1$ models on the third sub-sample, and the resulting preferred model is chosen on the basis of lowest AIC. Once the process is repeated for all sub-sample permutations, the final model is selected based on out-of-sample AIC.[3]

Perez-Amaral *et al.* (2003, p.4) state that there is no theoretical justification for the use of the two different evaluation measures (MSFE and AIC) except that the authors found the procedure to work well in simulations. AIC is useful when searching for a more parsimonious model as it penalises large models, which is important when there are a large number of candidate variables. Both criteria use disjoint sub-samples to obtain an 'out-of-sample' evaluation statistic, which avoids over-parameterisation because the use of disjoint sub-samples is a powerful mechanism for controlling the null rejection frequency, but at the expense of the non-null rejection frequency (see Hendry and Krolzig, 2004c).

[2]The λ varies in increments of approximately 0.1, resulting in $\nu \approx 9$.

[3]The criterion for selecting the final model depends on the econometrician's objectives. Perez-Amaral *et al.* (2005) state that the model which has the best performance over the whole sample is selected. Either AIC or cross MSFE (CMSFE) can be used, where CMSFE is computed by using two of the three sub-samples for estimation and the third for cross-validation. The overall CMSFE is obtained by summing the MSFEs from each of the three rotations.

The regressors in the candidate model depend on the order of inclusion of the variables, determined by absolute correlations with y. If the set of $\mathbf{W}$ regressors is larger than the number of observations in one sub-sample, the search over varying threshold parameters, λ, becomes essential. For low values of λ the program will work with many transformations, as only a small number will be selected. As λ gets closer to 1, only the $N_k - 1$ ranked variables (where N_k is the number of observations in sub-sample k) will be considered. Methods based on subset selection, as described in section 2.3.1 for PcGets, would be problematic as the rankings based on $|\widehat{\rho}_{kj}|$ apply to all candidate regressors and this ranking would be destroyed if subsets were considered. The program can be run using all level 1 transformations including those that are linearly dependent (note that the candidate model will only contain regressors that are linearly independent because of the collinearity index) ensuring that any transformation can potentially be included.[4]

2.3 PcGets

PcGets is a procedure for automatic model selection that is designed to select a parsimonious undominated representation of a GUM. The GUM is a very general model that includes all potentially relevant factors, and the *Gets* procedure tests downwards from this starting point to select a specific model. Krolzig and Hendry (2001) examine the properties of the model selection procedure in a linear framework: the LDGP can be found almost as often commencing from a general model as from the LDGP itself, and false rejection frequencies of null hypotheses can be controlled, correct rejections of alternatives are close to the theoretical upper bound and model selection is consistent. PcGets is described below (see Hendry and Krolzig, 2001, for details).

2.3.1 The PcGets selection algorithm

There are four stages in the PcGets algorithm, including estimation and testing of the GUM, the pre-search process, the multipath search procedure, and finally post-search

[4]RETINA contains a number of settings including: (i) selecting from level 0 regressors, providing a baseline linear model; (ii) selecting from all level 0 regressors and all level 1 transformations; and (iii) selecting from all level 0 regressors and all level 1 transformations excluding interaction terms, i.e. $\zeta(\mathbf{X}) = X_{i,h}^{\alpha} X_{i,h}^{\beta}$ for $\alpha, \beta = -1, 0, 1$. The new release of RETINA (see Marinucci, 2005) also enables the inclusion of regressors that interact with level 0 regressors or level 1 transformations, which is useful for interaction dummies.

evaluation. Initially, the GUM is formulated based on theory and previous evidence, which is sufficiently general to nest the LDGP. If there are relevant variables that are omitted from the GUM, this would be likely to manifest itself as a failure of the mis-specification tests. A batch of mis-specification tests are performed on the GUM to ensure congruency, which means that the model matches the unknown LDGP in all measured aspects (see Bontemps and Mizon, 2003). If there is a failure of congruency, the search procedure would not commence and the econometrician would need to re-specify the GUM. Observe that the LDGP is just the DGP for the set of variables under analysis, and so re-specifying the GUM implies that an alternative LDGP is being approximated. There are many different LDGPs that exist for varying sets of regressors, which could potentially all be congruent, implying that alternative congruent models can be obtained for different information sets. However, for some sets of variables, the targeted LDGP may be 'poor', and this is often detected by evidence of non-constancy.

Pre-search reduction tests are undertaken at loose significance levels to remove highly insignificant variables. Then the multipath search strategy commences from every feasible deletion, searching along each path using t-tests and F-tests until no more reductions can be made, checking the diagnostics at each reduction to ensure congruence. The resulting models are denoted terminal models, which are tested against their union until a unique, undominated, congruent model is selected.[5] Sub-sample reliability of the final model is evaluated by using overlapping sub-samples to formulate a reliability weighting, depending on whether a variable is significant in the two sub-samples and the full sample or not.

Various search strategies are available in PcGets, including the liberal strategy, the conservative strategy, the expert-users strategy and the quick modeller. The liberal strategy reduces the non-selection probability, whereas the conservative strategy is tighter, reducing the non-deletion probability. A conservative strategy is recommended if there are highly significant variables among many insignificant variables. If there are fewer regressors and the significant variables have smaller t-values, a liberal strategy is suggested.

[5]For the case in which a unique model does not emerge and the models are mutually encompassing and undominated, selection of the preferred model is made on the basis of information criteria.

Both are examined in the empirical examples below. The expert-users strategy allows for all selection strategy options to be set by the user; and the quick modeller lets the algorithm make all the modelling decisions, including selecting lag length, seasonals and outlier dummies, and transforming an I(1) model to a differenced, cointegrated model, upon which the model selection algorithm is then applied.

An important development is that PcGets can handle more variables, n, than observations, T. This enables a greater flexibility when examining nonlinear models, as the number of potential regressors is likely to be large. The GUM cannot be estimated initially. Instead, the variables are divided into J groups (we take $J = 2$ in this exposition but the analysis is easily generalised) in which the dimensions of the two groups are strictly smaller than T (and preferably $\leq \frac{T}{2}$). PcGets selects the first terminal model from the first subset of variables, and likewise the second terminal model from the second subset. The two terminal models are combined and used as the GUM in the second estimation stage by PcGets, resulting in a final terminal model. If the variables are not orthogonal, it is recommended that the variables are 'crossed-over'. This requires the two groups of variables, $\mathbf{x}_{1,t}$ and $\mathbf{x}_{2,t}$, to be partitioned into two halves, $\mathbf{x}^a_{i,t}$ and $\mathbf{x}^b_{i,t}$ for $i = 1, 2$, and cross-paired, resulting in six combinations, which are denoted the GUMs for the search procedure outlined above.[6]

We assert that under the null, for a correctly calibrated nominal significance level α, $n\alpha$ variables should still be retained on average even if $n > T$, as discussed below. To examine the actual properties of selection under the null when there are more variables than observations, we undertake a simple Monte Carlo experiment in which there are 40 irrelevant variables and 20 observations. The GUM is:

$$y_t = \sum_{i=1}^{40} \beta_i x_{i,t} + \epsilon_t, \qquad \epsilon_t \sim \mathsf{IN}\left[0, 1\right] \tag{2.3}$$

$$\mathbf{x}_t \sim \mathsf{IN}_{40}\left[\mathbf{0}, \mathbf{I}\right] \tag{2.4}$$

for $t = 1, ..., T$, where $T = 20$. The DGP is given by:

$$y_t = v_t, \qquad v_t \sim \mathsf{IN}\left[0, \sigma_v^2\right]. \tag{2.5}$$

[6]The 'crossed pairs' would include $\mathbf{x}^a_1\mathbf{x}^b_1$, $\mathbf{x}^a_1\mathbf{x}^a_2$, $\mathbf{x}^a_1\mathbf{x}^b_2$, $\mathbf{x}^b_1\mathbf{x}^a_2$, $\mathbf{x}^b_1\mathbf{x}^b_2$ and $\mathbf{x}^a_2\mathbf{x}^b_2$ for $J = 2$.

Table 2.1: Monte Carlo results: more variables than observations

	Average retained			*0 retained*			*1 retained*		
	Lib	Cons	Lib/Cons	Lib	Cons	Lib/Cons	Lib	Cons	Lib/Cons
MC	2.03	0.38	1.20	0.09	0.71	0.41	0.31	0.22	0.28
Theory	2.00	0.40	0.40	0.13	0.67	0.67	0.27	0.27	0.27

Notes: $n = 40$ variables; $T = 20$ observations; $M = 100$ replications.

The variables are divided into subsets of 10 and are selected using the liberal and conservative strategies. The union of the resulting terminal models is formed and selected again. We consider three cases: the liberal strategy for both stages, the conservative strategy for both stages, and the liberal strategy for the first stage and the conservative strategy for the second stage. Results are reported in Table 2.1.

The average number of variables retained in selection t-testing can be calculated using the binomial expansion of $1 = (\alpha + (1 - \alpha))^n$, and is given by:

$$\sum_{i=0}^{n} i \frac{n!}{i!\,(n-i)!} \alpha^i \,(1-\alpha)^{n-i} = n\alpha, \tag{2.6}$$

where n is the number of irrelevant variables and α is the significance level. Hence, with 40 variables, the liberal strategy should retain two variables and the conservative strategy should retain 0.4 of a variable on average. The Monte Carlo results closely match these theoretical results despite $n > T$. The probability that no variables are retained is $(1 - \alpha)^n$, and the probability of retaining one variable is $n\alpha\,(1 - \alpha)^{n-1}$, both of which are close to the Monte Carlo results, demonstrating that the selection properties of PcGets are unchanged with more variables than observations, even in the case of small samples. However, using the liberal strategy followed by the tighter conservative strategy results in a significance level of approximately 3%, which lies between the liberal and conservative strategies, suggesting that the significance levels used should be the same for each stage to ensure correct retention probabilities.

It is also worth noting the distinction between the costs of search and the costs of inference. The costs of search arise because the GUM is overly general and is tested down to find a congruent, undominated model. The costs of inference are unavoidable and are due to non-zero significance levels. They arise because the true specification is unknown and must therefore be tested for, even if the econometrician (unknowingly) starts with

the LDGP. Any selection strategy should be assessed on the basis of search costs alone as the costs of inference are inevitable, regardless of the selection procedure applied.

2.4 A comparison of the two approaches

Although RETINA and PcGets have many similar characteristics, a brief look at their selection algorithms reveals fundamentally different search strategies. A comparison of the two programs is provided in Perez-Amaral *et al.* (2005). This section briefly assesses the differences between the two algorithms, noting potential criticisms and counter arguments.

First, RETINA was developed with the aim of finding a model that has good out-of-sample predictive ability, whereas PcGets selects a congruent, dominant, in-sample model, aiming to locate the LDGP which is nested within the GUM. While in a stationary world, the preferred forecasting model is the best in-sample model, in a non-stationary world that is subject to structural breaks, this may not be the case. For time-series models in particular, a good forecasting model is often one that is robust to breaks (see Clements and Hendry, 1999). Nonlinear models are likely to be less robust than linear models, not only because estimation errors can accumulate rapidly, leading to forecast failure, but also because the interactions between variables are also likely to be subject to breaks. While RETINA does not insure against non-stationarities and structural breaks, the properties of RETINA as an "out-of-sample forecasting" tool are shown to be successful in cross-sectional applications, suggesting that RETINA may be more applicable to stationary data, and cross-section data in particular.

Specification of the nonlinear functions depends on the degree of interpretability that the modeller wishes to retain. PcGets specifies the GUM based on the econometrician's knowledge and experience, institutional knowledge, past evidence and economic theory. If the econometrician deems any nonlinear terms to be relevant, these should be included in the GUM. However, RETINA automates this decision and includes all transformations that have been specified by the program. A degree of economic interpretation is lost in RETINA, but as the goal is out-of-sample prediction, this property is not fundamental. The development of the PcGets 'Quick Modeller', which automatically selects the lag

length for the variables in the GUM, has advanced the automatic procedures embodied in PcGets, and generating nonlinear regressors as functions of the specified linear variables would be the next stage in automating the selection algorithm. As there are an infinite number of nonlinear transforms, an arbitrary cut-off point will be needed. Chapter 3 looks at classes of functions that might approximate economic data.

RETINA uses a specific-to-general approach, in which variables are added into the model sequentially, whereas PcGets uses a *Gets* approach that tests downwards from the GUM. There are numerous problems with the simple-to-general approach, discussed in Chapter 1. The specific-to-general methodology tends to have an *ad hoc* termination point for the search, and alternative path searches are unbounded, implying that the approach could miss the LDGP. Moreover, the null rejection frequency will not be controlled as the number of tests conducted will depend on the termination point, and failure of mis-specification tests is likely at the initial stages, invalidating conventional tests. The methodology also conflicts with the ideas behind encompassing. RETINA overcomes these problems by constructing the GUM as a starting point, effectively bounding the number of paths, and by using a grid search over threshold values as the termination point for each path search. RETINA also avoids testing, thereby solving the problems of mis-specification and uncontrolled test size. This does mean that there is no guarantee that the final model selected by RETINA is congruent, which may or may not be relevant for forecasting models.

PcGets performs an exhaustive search ensuring that almost all paths are checked, whereas RETINA uses a selective search determined by correlations. This is particularly relevant when there are more potential regressors than observations, as the candidate models will depend on the order of inclusion of variables. RETINA does control for collinearity through the use of the threshold parameter. An orthogonal specification of the GUM is preferable in PcGets, as collinearity induces a fall in the non-null rejection frequency and an increase in the null rejection frequency (see Hendry and Krolzig, 2005, for Monte Carlo evidence on collinearity). For forecasting time-series, controlling for collinearity will be essential if there are interactions between collinearity and structural

breaks (see Hendry and Hubrich, 2006).

RETINA uses three disjoint sub-samples to cross-validate the variables selected. The PcGets approach performs a post-selection reliability check using sub-samples to evaluate whether significance is substantive or due to chance. There is a null rejection frequency gain in the RETINA procedure due to the use of sub-samples. However, Lynch and Vital-Ahuja (2003) show that the use of sub-samples for model selection delivers no gain over using tighter significance levels for the full sample. This is because there is a trade-off between the null and non-null rejection frequencies. The non-null rejection frequency function is highly nonlinear and depends on the LDGP. Thus, the key question is whether a null rejection frequency reduction for the whole sample, which is equivalent to a null rejection frequency gain caused by the use of sub-samples, results in an increase or decrease in the non-null rejection frequency. Hendry and Krolzig (2004c) examine the case of non-overlapping sub-samples. With J equal partitions, defining t_0 as the full sample t-value and t_j as the jth sub-sample t-value, the full sample t-value can be calculated as a function of the sub-sample t-values. If τ_j is the fraction of observations in sub-sample j (where $\sum_j \tau_j = 1$), Hendry and Krolzig show that:

$$\mathsf{t}_0 \simeq \sum_{j=1}^{J} \sqrt{\tau_j}\mathsf{t}_j > \sum_{j=1}^{J} \tau_j \mathsf{t}_j. \tag{2.7}$$

The weighted sum of the sub-sample t-values is less then the full sample t-value. If $J = 3$, as in RETINA:

$$\mathsf{t}_0 \simeq \frac{1}{\sqrt{3}} \sum_{j=1}^{3} \mathsf{t}_j = 0.58 \sum_{j=1}^{3} \mathsf{t}_j. \tag{2.8}$$

Hence, selecting on the sub-sample t-values, a marginally significant sub-sample t-value of 2 will imply a full sample t-value of 3.5. A direct relationship between the average t^2-value on the full sample and sub-samples is given by:

$$\mathsf{E}\left[\mathsf{t}_j^2\right] \simeq \frac{1}{J}\psi^2 \tag{2.9}$$

where ψ is the non-centrality parameter. There is a reduction in the information content of the sub-sample t-tests, which Hendry and Krolzig (2004c) refer to as the 'curse of sub-samples', implying that locating the LDGP will be more difficult using sub-sample

procedures. For non-stationary data, if the breakpoints are known, substantial gains could be achieved by analysing sub-sample information. Conversely, the best in-sample fit in one sub-sample may well differ from that in a second sub-sample because of structural breaks. In this case, the use of cross-validation would be more appropriate for post-sample evaluation as opposed to informing model selection.

While PcGets controls the rejection frequency of the test battery, RETINA does not address the question of costs of search and costs of inference. Krolzig and Hendry (2001) and Hendry and Krolzig (2003b) examine the properties of PcGets and show that, even for small samples, the rejection frequencies of the t-tests are close to their theoretical levels based on a t-test from the true model. Distinguishing between individual significance levels, η, and the overall significance level of a test battery of n mis-specification tests, the overall rejection probability under the null is given by $1-(1-\eta)^n$, which can be controlled. RETINA selects variables on the basis of out-of-sample MSFE rather than statistical tests, reflecting the differing objectives of the two procedures.

2.5 Applications

This section assesses evidence on RETINA and PcGets. First, the results in Perez-Amaral *et al.* (2005) (hereafter referred to as PGW) are summarised and re-assessed in light of some important considerations, including the presence of substantial outliers, the arbitrary re-scaling of the data, the assessment of forecast ability based on in-sample rather than out-of-sample evaluation, and the selection algorithm for nonlinear functions. Secondly, we assess a time-series application using data from Hoover and Perez (1999) (hereafter referred to as HP), evaluating both algorithms' ability to recover the known LDGP under the null of linearity.

2.5.1 Cross-section data: Demand for business toll telephone services

PGW assess the two programs using data from a cross-section of US firms regarding the demand for business toll telephone services in 1997. The authors model the duration of intra-lata (within local access and transport area) calls (denoted $INTRA$) with the explanatory variables including the number of business lines (bus), hunting lines (hun),

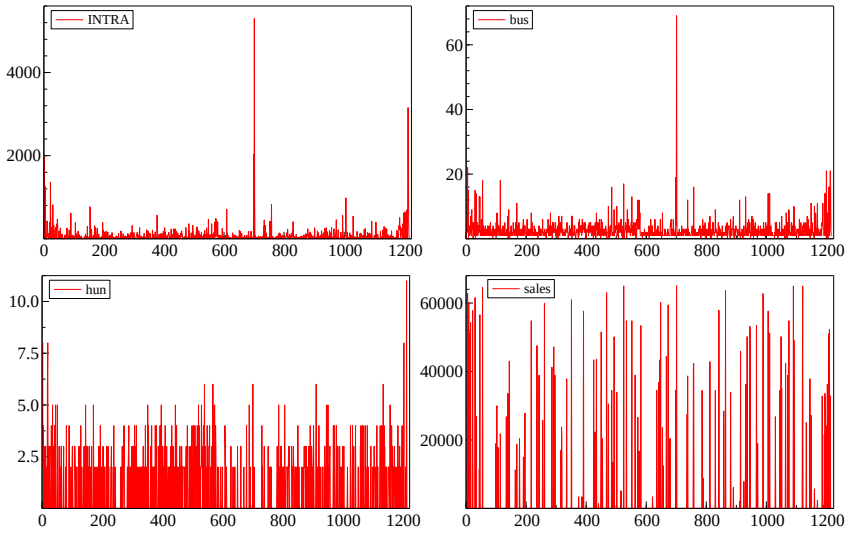

Figure 2.1: Cross-sectional data including $INTRA$, *bus*, *hun* and *sales*.

sales of the company (*sales*), number of employees working locally (*emt*), total number of employees for the business (*emh*), physical size of the business proxied by square feet of the premises ($sqft$) and population of the business area location (*pop*). The sample size is 1,217 and the data are initially re-scaled to avoid large differences in the magnitudes of the variables. Observations are randomly assigned to the three sub-samples and the models are assessed on the basis of two criteria: AIC and CMSFE. The data are recorded in Figures 2.1 and 2.2.

Correcting for outliers

There are substantial outliers in the data, and the nonlinear terms retained by PGW primarily capture these outliers by proxying dummies, rather than by modelling the nonlinear characteristics of the data. Using a standard method of outlier removal in which all observations of $INTRA$ outside the range $(\widehat{\mu} \pm 2\widehat{\sigma})$ are removed, 22 observations would be excluded, resulting in a substantial reduction in variance. However, this crude method of outlier removal has many problems. The standard error is dependent on the outliers and so the rule determining exclusion is a function of the excluded variables, and the method removes extreme observations in the dependent variable rather than extreme observations in the residuals. This may remove the underlying nonlinearity instead of removing just the outliers.

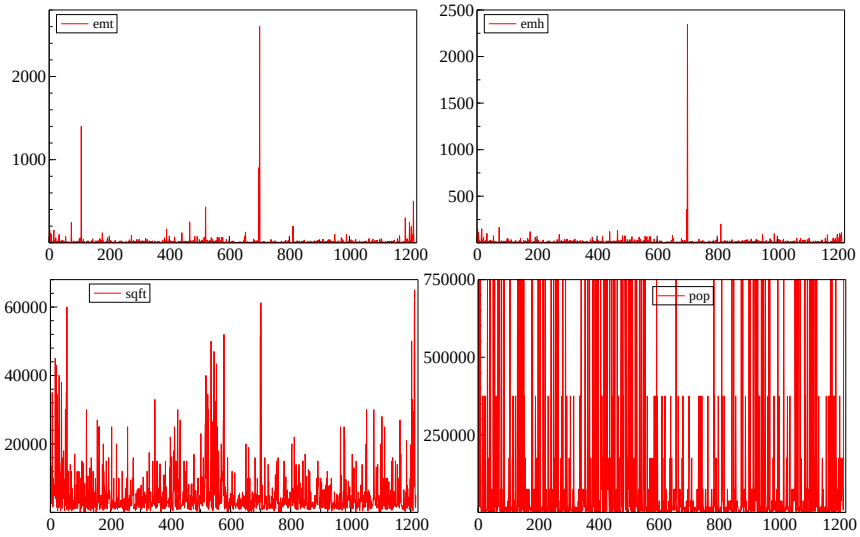

Figure 2.2: Cross-sectional data including *emt*, *emh*, $sqft$ and *pop*.

Hendry *et al.* (2004) have proposed the use of saturation with indicator variables to detect outliers, utilising the fact that PcGets can handle more variables than observations. The testing procedure initially generates indicator variables for every observation, $I_t = 1_{\{t=t_i\}} \forall T$. A regression of y on all indicators would result in a perfect fit. Instead, the indicators are divided into J equal subsets which form the initial GUMs and PcGets selects the significant indicators from each subset, which are then stored.[7] The joint model is formulated as the union of the terminal models and PcGets re-selects the indicators. Under the null that there are no outliers, αT indicators will be retained on average for a significance level α.

If there are n potentially relevant variables and T indicators, then $n+T >> T$. Using the technique in which the variables are divided into subsets, n_k for $k = 1, ..., K$ and T_j for $j = 1, ..., J$, we can formulate JK GUMs in which all cross pairings of $n_k + T_j$ $(\forall j, k)$ are included. The same selection procedure is applied, but by selecting the outliers and the relevant variables at the same time the indicator saturation procedure is effectively applied to the residuals of the dependent variable as opposed to the variable itself, thereby avoiding spurious retention of indicators. The process can be performed iteratively if the union after the first reduction stage is still larger than T.

The nonlinear general model consists of 86 explanatory variables. These include an

[7] An intercept is also included in the GUMs.

intercept, the seven explanatory variables and all level 1 transformations, excluding any transformations that result in numerical problems. As some observations of *bus* and *hun* are 0, inverse transformations of these explanatory variables are excluded. We generate $J = 6$ subsets of indicators and specify the six GUMs (with 288 or 289 regressors), which include an intercept, all explanatory variables and $\frac{T}{J}$ indicator variables. A conservative strategy is used given the large number of regressors, and the block F-test is adjusted to avoid missing a significant variable among a plethora of insignificant indicators. Hence, the final selected model should indicate the outliers at the 1% significance level. Forty-six indicators are retained ($< 4\%$ of observations), with an average t-value of 5.7. The largest t-value is 17.6, indicating that there are substantial outliers in the data. The extreme observations that remain correspond to nonlinearities in the data that can be explained by the explanatory variables, providing evidence for nonlinearity. The removal of outliers has a significant effect on the models retained; there is a dramatic reduction in the forecast error and fewer nonlinear regressors are retained than in PGW.

As a benchmark, we examine the linear model in which just the seven explanatory variables and an intercept are included in the general model. Applying the indicator saturation test to this model, using $J = 3$ subsets of indicators, 27 indicators are retained in the final model, of which the average t-value is 8.9 and the highest t-value is 34. The inclusion of the nonlinear functions reduces the significance of the outliers on average, but increases the number of outliers identified. Almost all of the outliers identified in the linear model are common to the nonlinear model as well.

Variable transformation

PGW initially re-scale the data as there are large differences in the scales of the variables. The scales are given in (2.10). The problem with this arbitrary re-scaling is that when forming the matrix of regressors, $\mathbf{X}$, using all level 1 transformations, the diagonal elements of the $\mathbf{X}'\mathbf{X}$ matrix are still very different. One solution would be to take a log-transformation to re-scale the data. This nonlinear transformation has the benefit of downweighting the largest observations which are dominating other features. As *bus* and *hun* both take the value of 0 for some observations, the transformation $\ln(x + 1)$ is used

for these two variables. The problem with the log-transformation is that the forecasts of $\ln(INTRA)$ would have to be transformed back to levels and this will generate some sub-optimality. In the analysis below, both levels and logs data are considered.

$$\begin{bmatrix} INTRA & \div 10 \\ bus & \text{No change} \\ hun & \times 10 \\ sales & \div 1,000,000 \\ emt & \div 10 \\ emh & \div 10 \\ sqft & \div 1,000 \\ pop & \div 100,000 \end{bmatrix}. \tag{2.10}$$

The data are randomly re-ordered to ensure homogeneous subsets. A plot of the data provides evidence of heteroskedasticity, with clusters of observations with a larger variance at the beginning and end of the sample, and the random re-ordering of the data removes any systematic component that may be causing heteroskedasticity.

Table 2.2 records descriptive statistics of both the levels and log-transformed data, based on the randomly re-ordered data set. In order to assess the forecasting performance of the models, the final quarter of the observations are excluded from the outlier analysis. These observations are retained to ensure an out-of-sample forecasting analysis. The indicator saturation tests are performed over a sample of 913 observations. For the levels data, 21 outliers are identified for the linear model and 31 for the nonlinear model. Sixteen of these outliers are common to both models. Taking a log-transformation results in fewer outliers; 11 in the linear model and 13 in the nonlinear model, eight of which are common to both models, implying that few are due to the nonlinear functions. Removing outliers based on the $\widehat{\mu} \pm 2\widehat{\sigma}$ rule would have led to 65 observations being removed for the log-transformed data, which is a much more stringent strategy.

Forecasting out-of-sample

When using the full sample to derive a model, the forecast errors generated from the model will be biased downwards because information in the forecast period will be included in the model. Instead, a forecast evaluation is required in which the forecasts are assessed on out-of-sample data. Hence, analogous to the indicator saturation tests,

Table 2.2: Descriptive statistics of the dependent variable, $INTRA$

	Mean		*Standard Deviation*	
	In-sample	*Forecast*	*In-sample*	*Forecast*
$INTRA$	8.455	7.803	25.461	24.032
$INTRA$ exc. linear outliers (21)	6.722	7.803	20.613	24.032
$INTRA$ exc. nonlinear outliers (31)	6.551	7.803	22.695	24.032
$\ln(INTRA)$	3.179	3.107	1.667	1.764
$\ln(INTRA)$ exc. linear outliers (11)	3.220	3.107	1.618	1.764
$\ln(INTRA)$ exc. nonlinear outliers (13)	3.226	3.107	1.615	1.764

Notes: (.) denotes number of outliers.

the models are selected using the first 913 observations. Rather than assess the models on CMSFE as in PGW, we use root mean square forecast error (RMSFE) and mean absolute percentage error (MAPE):

$$RMSFE = \sqrt{\frac{1}{H}\sum_{i=N+1}^{N+H}(y_i - \widehat{y}_i)^2} \tag{2.11}$$

$$MAPE = \frac{100}{H}\sum_{i=N+1}^{N+H}\left|\frac{y_i - \widehat{y}_i}{y_i}\right| \tag{2.12}$$

These statistics are computed for the final 304 observations. To assess forecast accuracy relative to the in-sample model fit, we can compare the equation standard error augmented by the standard deviation of the indicators to the RMSFE.

To determine whether the forecasts are statistically different, we use a test based on the Morgan-Granger-Newbold (MGN) test of comparative forecast accuracy (see Clements and Hendry, 1998b, p.323).[8] As this test is oversized if the errors are not normally distributed, we use a modified test developed by Harvey *et al.* (1997). Defining the forecast errors of two competing forecasts at $N+h$ as $\widehat{e}_{N+h}$ and $\widetilde{e}_{N+h}$, we can undertake an orthogonalising transformation given by:

$$u_{1,N+h} = \widehat{e}_{N+h} - \widetilde{e}_{N+h} \tag{2.13}$$

$$u_{2,N+h} = \widehat{e}_{N+h} + \widetilde{e}_{N+h} \tag{2.14}$$

Setting this in a regression framework, a test for forecast differences is equivalent to a test of zero correlation between $u_{1,N+h}$ and $u_{2,N+h}$, which can be tested with the hypothesis

[8]The MGN test is a test of equality of MSFE, which is equivalent to equality of variances if we make the assumption that the forecast errors are unbiased. Other assumptions are that the forecast errors are normally distributed and serially uncorrelated.

$H_0 : \beta = 0$ in the regression:

$$u_{2,N+h} = \beta u_{1,N+h} + \epsilon_{N+h}, \quad \text{for } h = 1, ..., H. \tag{2.15}$$

The test statistic is:

$$S = \widehat{\beta} \left[\frac{\sum_{i=N+1}^{N+H} u_{1i}^2 \widehat{\epsilon}_i^2}{\left(\sum_{i=N+1}^{N+H} u_{1i}^2 \right)^2} \right]^{-\frac{1}{2}} \underset{app}{\sim} \mathsf{t}_{H-1}(0) \tag{2.16}$$

The null distribution is not exact but the test is shown to have good properties in large samples.[9] The test enables two competing forecasts of the same phenomena to be judged as to whether the seemingly preferable forecast based on MSFE (which is equivalent to minimum variance under the assumption of unbiased forecasts) is a better forecast, or whether such an outcome is due to chance. Rejection of the null of equal forecast accuracy implies the forecast with the lower MSFE is a statistically superior forecast, based on the MSFE criterion.

Strategies for searching over nonlinear functions

Various strategies need to be considered if many nonlinear functions are included in the GUM. Retaining many nonlinear functions in the model is likely to be detrimental to forecasting as the model will lose robustness. However, for linear models, Hendry and Hubrich (2006) find that variables with even relatively insignificant coefficient estimates should contribute to forecasting, although this is derived from the theory of predictability and therefore abstracts from many of the problems associated with forecastability.

Given their large number, it is important to narrow down the number of nonlinear regressors. The ideal strategy would be to use more stringent significance levels for the nonlinear functions compared to the linear functions, testing both sets of regressors within the same procedure. Two methods that are used in the application include pre-searching over nonlinear functions and discarding subsets of nonlinear functions based on block F-tests.

A pre-search over the nonlinear functions is conducted by regressing the nonlinear

[9] $H = 304$ in this application.

functions on an intercept and the linear functions, retaining the residuals: an application of the Frisch and Waugh (1933) theorem. PcGets is then used to test the significance of these residuals using a stringent search strategy (significance level of the individual t-tests of 0.005). The selected residuals are replaced by their nonlinear functions in the GUM, which also contains all linear functions, and testing is conducted at standard significance levels. This technique is applied to the levels data, ensuring a tighter selection criterion for the nonlinear functions. Without applying the pre-search strategy, 25 regressors are retained for the liberal strategy. This is reduced to 11 when using a significance level of $\alpha = 0.005$ for the nonlinear functions, with just four nonlinear functions being retained, all of which have $|t|s > 10$.

The level 1 transformations include various classes of functions including squares, inverses, cross-products, etc. The number of regressors in the GUM may be narrowed down by undertaking block F-tests on these classes of functions, deleting any insignificant sets of regressors and thereby reducing the number of variables that the algorithm needs to search over. Block F-tests are computed for the log-transformed data, testing various subsets of the GUM, and a sequential procedure is followed until the classes of nonlinear functions that are significant are retained. These then form the GUM, along with the linear functions, and standard testing procedures are applied.

Results

Table 2.3 records PGWs results for comparison.[10] A linear GUM, consisting of an intercept and the seven explanatory variables, is examined. Four models are computed, including a linear regression with no selection (denoted bench), two PcGets models using the liberal and conservative strategies respectively, and RETINA with just the level 0 variables. The GUM for the nonlinear models includes all level 1 transforms excluding any regressors with inverses of 0, resulting in 86 variables in the GUM. The results show that, for the linear models, RETINA has a better forecast performance, whereas PcGets performs better for the nonlinear models, which is remarkable as RETINA is designed explicitly to select nonlinear forecasting models whereas PcGets is currently designed to

[10] Parameter estimates are not reported, see PGW for a full outline of the results.

Table 2.3: Perez-Amaral *et al.* (2005) results for cross-sectional data.

	Linear Models				*Nonlinear Models*			
	Linear	*PcGets*		*RETINA*	*PcGets*		*RETINA*	
	bench	*Lib.*	*Cons.*	*level 0*	*Lib.*	*Cons.*	*AIC*	*CMSFE*
CMSFE	909.88	896.11	903.08	770.86	507.42	**498.63**	572.01	518.00
AIC	5.443	5.440	5.446	5.459	**4.839**	**4.839**	4.932	4.947
$\overline{R^2}$	0.603	0.603	0.600	0.595	0.785	0.784	0.757	0.756
Parameters	8	6	5	5	19	18	9	9

Notes: Lowest CMSFE and AIC are highlighted in **bold**.

select linear models. Furthermore, non-parsimony is not detrimental to forecasting, as PcGets has an extra 9 or 10 parameters in the nonlinear model compared to RETINA, and yet its forecasting performance is superior.

The results for the re-scaled data given by (2.10) (denoted levels) are reported in Table 2.4 and the results for the log-transformed data are reported in Table 2.5. Examining the levels models, which are comparable to PGW, a notable feature of the results is that parsimony is obtained by removing outliers. The fit is also improved. The nonlinear models selected by PcGets have the highest $\overline{R^2}$ and the lowest AIC, suggesting that they have the best in-sample fit. Only one nonlinear variable that is retained is the same as in PGW. The t-statistics for the nonlinear functions retained by PcGets are all > 10, and yet RETINA does not retain any of these variables. Conversely, the variable that RETINA does retain has a t-value of 30, which is not retained by PcGets. The explanation for these results is collinearity between those in PcGets and that in RETINA, and arises from RETINA's selection criterion.

In terms of forecasting performance, the nonlinear models selected by PcGets have the lowest MAPE and RMSFE, suggesting that the forecasting performance of these models is superior. A test of forecast error equivalence is reported in Table 2.6, in which the Harvey *et al.* (1997) statistic is recorded in the lower diagonal for the levels model. The results show that none of the forecasts are statistically different from each other when evaluated in terms of MSFE. However, all other forecasts lie outside the ± 2 standard error bands of the MAPE of the PcGets nonlinear models, suggesting that these forecasts are superior. RMSFE is not unit-free or robust to extreme observations, implying that comparisons of forecasts based on RMSFE are difficult. The results do,

Table 2.4: Models selected for the demand for business toll telephone services

Levels	*Linear Models*				*Nonlinear Models*		
	Linear	*PcGets*		*RETINA*	*PcGets*		*RETINA*
	benchmark	*Lib.*	*Cons.*	*level 0*	*Lib.*	*Cons.*	*level 1*
RMSFE	14.00	14.05	14.20	14.23	13.74	**13.70**	14.54
MAPE	1084.2	1016.8	942.13	1020.4	766.42	**752.45**	1169.7
AIC	4.521	4.515	4.518	4.528	4.298	4.301	4.561
$\overline{R^2}$	0.786	0.786	0.785	0.783	0.859	0.858	0.815
Parameters	8	4	3	3	11	10	3
$Intercept$	−0.161 (0.555)			0.991 (0.464)	3.190 (0.551)	3.303 (0.550)	3.991 (0.333)
Bus	1.182 (0.174)	1.216 (0.129)	1.305 (0.123)	1.300 (0.152)	2.238 (0.152)	2.098 (0.137)	
Hun	0.067 (0.024)	0.065 (0.022)	0.077 (0.021)		−0.205 (0.026)	−0.192 (0.025)	
$Sales$	4.099 (34.041)				−67.85 (31.46)		
Emt	0.022 (0.067)				−0.482 (0.086)	−0.482 (0.086)	
Emh	1.864 (0.097)	1.880 (0.056)	1.884 (0.056)	1.883 (0.060)	−0.865 (0.242)	−0.933 (0.241)	−0.992 (0.119)
$Sqft$	0.092 (0.046)	0.096 (0.042)			−0.752 (0.100)	−0.764 (0.100)	
Pop	0.210 (0.137)						
$(emh)^2$					−0.594 (0.044)	−0.588 (0.044)	
$(sqft)^2$					0.032 (0.003)	0.032 (0.003)	
$hun \times emh$					0.155 (0.008)	0.156 (0.008)	
$emt \times emh$					0.546 (0.040)	0.541 (0.040)	
$bus \times sqft$							0.184 (0.006)

Notes: Data are in levels, scaled using the rules in equation (2.10). Evaluation statistics are reported in rows 1-5 with the best forecasting performance highlighted in **bold**; coefficient estimates are reported in the remaining rows with standard errors in parentheses.

Table 2.5: Log-transformed models for demand for business toll telephone services

Logs	*Linear Models*				*Nonlinear Models*		
	Linear	*PcGets*		*RETINA*	*PcGets*		*RETINA*
	benchmark	*Lib.*	*Cons.*	*level 0*	*Lib.*	*Cons.*	*level 1*
RMSFE	1.643	1.649	1.649	1.645	**1.642**	1.649	1.644
MAPE	185.63	186.66	186.66	**183.63**	186.63	187.48	186.54
AIC	0.774	0.768	0.768	0.779	0.742	0.746	0.761
$\overline{R^2}$	0.179	0.179	0.179	0.170	0.200	0.195	0.183
Parameters	8	3	3	3	5	4	3
Intercept	0.333 (0.449)			1.693 (0.122)		1.636 (0.120)	2.196 (0.101)
Bus	0.791 (0.109)	0.894 (0.097)	0.894 (0.097)	0.877 (0.099)		1.373 (0.107)	
Hun	0.102 (0.087)						
Sales	0.035 (0.026)						
Emt	0.089 (0.165)			0.288 (0.052)			
Emh	0.125 (0.172)	0.219 (0.056)	0.219 (0.056)				0.284 (0.056)
Sqft	0.163 (0.051)	0.224 (0.016)	0.224 (0.016)		0.208 (0.016)		
Pop	0.014 (0.025)						
$(bus)^2$							0.328 (0.036)
$bus \times hun$					−0.371 (0.110)	−0.610 (0.097)	
$hun \times emh$					0.256 (0.064)	0.412 (0.052)	
$emt \times sqft$					0.020 (0.006)		
$bus \div sqft$					8.657 (0.927)		

Notes: Data are log-transformed. Evaluation statistics are reported in rows 1-5 with the best forecasting performance highlighted in **bold**; coefficient estimates are reported in the remaining rows, with standard errors in parentheses.

Table 2.6: Harvey *et al.* (1997) test statistics of equivalent forecast errors

	Linear bench	*Linear Lib.*	*Linear Cons.*	*Linear RETINA*	*Nonlin Lib.*	*Nonlin Cons.*	*Nonlin RETINA*
Linear bench	-	-0.619	-0.731	-0.209	0.030	-0.366	-0.036
Linear Lib.	-0.662	-	-0.627	0.208	0.318	-0.124	0.304
Linear Cons.	-0.756	-1.273	-	0.240	0.337	-0.105	0.329
Linear RETINA	-0.781	-1.535	-0.171	-	0.184	-0.282	0.174
Nonlin Lib.	0.719	0.733	0.788	0.808	-	-0.725	-0.063
Nonlin Cons.	0.770	0.762	0.815	0.836	0.453	-	0.390
Nonlin RETINA	-0.093	0.096	0.136	0.929	-0.360	-0.372	-

Notes: Levels results are reported below the diagonal and log-transformed results are reported above the diagonal.

however, suggest that the PcGets selection procedure with a pre-search over nonlinear functions yields 'good' forecasts.

For the log-transformed results, even more parsimony is attained in PcGets. Block F-tests on the classes of functions removed both the squared-inverses and cross-inverses from the GUM. Again, the nonlinear PcGets liberal model has the best in-sample fit based on $\overline{R^2}$, with the lowest AIC. The linear RETINA model produces the best forecasts based on MAPE and the PcGets liberal strategy performs best on RMSFE. The forecasts are very comparable and it is difficult to rank the models definitively. The upper diagonal of Table 2.6 records the forecast comparison statistics, none of which are significant. Only one of the nonlinear variables is the same for PcGets in the levels and logs models, with none for RETINA, indicating there are no clear relevant variables and various nonlinear functions will proxy the underlying characteristics of the data.

There is a vast improvement from removing outliers and ensuring a homogeneous data sample. In addition, forecast accuracy must be conducted on out-of-sample forecasts or the results will be biased. Finally, we advocate stringent critical values for nonlinear functions as these can drive forecasts badly awry, but looser critical values are required for the linear functions. Overall, the results improve on those presented in PGW, although the forecast errors are all quite similar. However, the null rejection frequency costs of searching for a nonlinear model are small for RETINA, suggesting that if it is unknown whether the LDGP contains nonlinearities and interaction terms, the use of RETINA could be informative.

2.5.2 Time-series data: Hoover and Perez (1999)

To assess the programs on time-series data, we apply both RETINA and PcGets to a drawing from the Hoover and Perez (1999) data set. The data are outlined in Table 2.11 in the Appendix. We examine four of the DGP specifications outlined in HP, which are given in equations (2.17) to (2.20).

$$y_{1,t} = 130u_t \tag{2.17}$$

$$y_{2,t} = 130u_t^* \tag{2.18}$$

$$y_{7,t} = 1.33\Delta FM1DQ_t + 9.73u_t^* \tag{2.19}$$

$$y_{9,t} = 0.67\Delta FM1DQ_t - 0.023\Delta^2 GGEQ_t + 4.92u_t^* \tag{2.20}$$

where:

$$u_t \sim \mathsf{N}[0,1] \tag{2.21}$$

$$u_t^* = 0.75u_{t-1}^* + u_t\sqrt{7/4}. \tag{2.22}$$

The linear GUM for RETINA and PcGets includes an intercept, variables dated t and $t-1$ of $\Delta DCOINC$, $\Delta^2 GD$, $\Delta^2 GGEQ$, $\Delta GGFEQ$, $\Delta^2 GGFR$, $\Delta GNPQ$, $\Delta GYDQ$, $\Delta GPIQ$, $\Delta^2 FMRRA$, $\Delta^2 FMBASE$, $\Delta FM1DQ$, $\Delta FM2DQ$, $\Delta FSDJ$, $\Delta FYAAAC$, ΔLHC, $\Delta LHUR$, ΔMU, $\Delta^2 MO$, ΔGCQ, and the lagged dependent variable, giving a total of 40 regressors. The sample is 1960q2-1995q1, and 20 observations are retained for an out-of-sample forecast assessment, giving an in-sample period of 1960q2-1990q1 ($T = 120$ observations). The nonlinear GUM for RETINA and PcGets includes the linear GUM as outlined and the level 1 transformations given by the squares, inverses, and inverse-squares. The interaction terms including cross-products, cross-inverses and ratios are excluded because of the limited number of observations. In total there are 157 regressors in the nonlinear GUM, including the nonlinear transforms of the lagged dependent variable, but we exclude the squared-inverses of $\Delta DCOINC$, $\Delta^2 GD$, $\Delta GYDQ$, $\Delta GGFEQ$, $\Delta^2 FMBASE$, $\Delta FM2DQ$, ΔLHC and $\Delta LHUR$, because they all have observations that exceed $2.5e^9$ causing numerical estimation problems. This results in 141 regressors. For the PcGets model selection, we partition the regressors into four groups and all pairwise combinations of the groups are implemented as GUMs, resulting in six combinations, as outlined in section 2.3.1. The specific models are then combined to produce the final GUM.

A summary of the results is reported in Table 2.7, with full results reported in Tables 2.12 to 2.15 in the Appendix. 'Relevant' refers to the number of relevant variables retained and 'Irrelevant' refers to the number of irrelevant variables that are retained. Note that

Table 2.7: Number of relevant and irrelevant variables retained for the DGPs based on Hoover and Perez (1999)

	Linear Models			*Nonlinear Models*		
	PcGets		*RETINA*	*PcGets*		*RETINA*
	Lib.	*Cons.*		*Lib.*	*Cons.*	
$y_{1,t}$ [0]						
Relevant	-	-	-	-	-	-
Irrelevant	4	3	2	12	4	2
$y_{2,t}$ [0]						
Relevant	-	-	-	-	-	-
Irrelevant	3	1	2	9	1	2
$y_{7,t}$ [1]						
Relevant	1	1	1	1	1	1
Irrelevant	3	0	1	7	3	1
$y_{9,t}$ [2]						
Relevant	1	1	1	1	1	1
Irrelevant	3	0	1	8	2	1

Notes: [.] denotes the number of variables in the DGP.

RETINA always retains the intercept and therefore it will always retain one irrelevant variable given the DGP specifications. RETINA tends to deliver a more parsimonious model than the PcGets liberal strategy when the nonlinear transformations are included in the GUM. The RETINA models chosen when the nonlinear functions are included are identical to those chosen when just the linear model is examined, indicating that the cost of testing for nonlinearity is low in RETINA if the DGP is linear. The sub-sample cross-validation procedure is analogous to imposing very stringent critical values, resulting in more parsimonious models. Given the DGPs in HP, this parsimony is advantageous. The linear PcGets model selected using the conservative strategy identifies the DGP for model 7. Interestingly, neither algorithm retains $\Delta^2 GGEQ_t$, which is a relevant variable in model 9.

The PcGets liberal strategy tends to overfit when nonlinear functions are included in the GUM. The strategy implemented should be adapted to the problem under consideration, and hence a conservative or more stringent strategy would be recommended when there are many nonlinear functions that have a high probability of being irrelevant. While the nonlinear models selected using the conservative strategy retain slightly more regressors than in the linear case, the costs of searching for nonlinear functions are not too high and can be controlled in PcGets. It is essential that the data pass diagnostics as

nonlinear terms may be kept to ensure congruence, even if they are insignificant. Overall, few nonlinear terms were retained.

RETINA and PcGets are similar with regard to forecast accuracy. The overfitted nonlinear model selected by the PcGets liberal strategy has the lowest RMSFE for model 1 but it has a substantially larger MAPE compared to the other models. This is due to the retained nonlinear functions generating substantial forecast outliers which are magnified by the MAPE criterion. For model 2, RETINA performs best based on RMSFE criterion but the PcGets conservative strategy outperforms RETINA on MAPE. The forecasts from both models are similar and it is difficult to draw substantive conclusions from the results. The models identifying the DGP have the lowest RMSFEs for model 7 (the intercepts in the RETINA models are negligible) but observe that the overfitted nonlinear model beats the DGP when assessing the forecasts on MAPE. This indicates that parsimonious models are not always preferable. A similar result is observed in model 9, with the nonlinear model chosen using the liberal strategy reporting the lowest MAPE.

To conclude, the results of both PcGets and RETINA are promising. RETINA tends to underfit, which may be useful for forecasting purposes as parsimonious models may be more robust. This is not necessarily the case as a less parsimonious model could be made robust, for example, by differencing. Simplicity in itself is not a necessary element of forecasting models but robustness often is. RETINA's more parsimonious criteria is driven by the use of disjoint sub-samples and the search strategy stage in which variables from the local best model are sequentially added in and chosen on the basis of AIC, although note that AIC does overselect asymptotically. The costs of search for PcGets are given in (2.6): for the linear GUM ($n = 40$) two variables will be retained on average at the 5% significance level but this increases to seven variables for the nonlinear GUM ($n = 141$). Campos *et al.* (2003) observe the need for significance levels to vary with both sample size and the number of candidate variables.

2.6 Monte Carlo evidence on RETINA and PcGets

This section provides a formal evaluation of the properties of nonlinear model selection, measuring the retention and deletion probabilities of RETINA and PcGets. One concern is that the inclusion of nonlinear regressors in the GUM may result in over-parameterisation of the model. This could occur due to commencing with a large GUM, in which case tighter significance levels should be used, or it could be driven by 'inference fragility', whereby the model is sensitive to keeping one variable which may have low significance. Once the marginal variable is removed, the significance of other correlated variables tends to fall. Other problems with nonlinearity include a potentially high level of collinearity between variables and their corresponding nonlinear transformations, and the presence of outliers whereby nonlinear functions proxy indicator dummies to capture these outliers.

The design of the Monte Carlo experiments is kept simple for tractability. There are a multitude of nonlinear and interaction terms that could be examined but the experiments concentrate on those generated by RETINA. We assess an orthogonal model in which the regressors are generated as white-noise processes, and a non-orthogonal model in which the regressors are independent normal innovation processes with a mean of ten.

Non-existence of moments is a concern with the orthogonal model. Nonlinear transformations such as the inverse or squared-inverse can explode with a zero-mean process. Observe that in Monte Carlo experiments, the existence of moments can be crucial. By increasing the number of replications, the probability that a draw will take a value very near zero is increased. For a small number of replications, the probability that a draw is zero is negligible, but this increases with the number of replications M. Hence, there is a dichotomy between standard Monte Carlo theory that requires an increase in the number of replications to determine asymptotic results and the problem of increasing the probability of a zero draw: see Sargan (1982). The non-orthogonal model ensures the existence of moments, but this model generates substantial collinearity between the linear and nonlinear functions.

2.6.1 Monte Carlo evidence on the orthogonal model

To assess the performance of PcGets and RETINA on orthogonal data, providing a baseline case, we examine a white-noise DGP based on Krolzig and Hendry (2001). The orthogonality of regressors enables population t-values to be easily calculated. The DGP is given by:

$$\begin{aligned} y_t &= \sum_{k=1}^{5} \beta_{k,0} x_{k,t} + \beta_{6,0} x_{1,t}^2 + \beta_{7,0} x_{2,t} x_{3,t} + \beta_{8,0} x_{4,t} x_{5,t} + \varepsilon_t, \qquad (2.23) \\ \varepsilon_t &\sim \mathsf{IN}\left[0,1\right] \end{aligned}$$

$$\mathbf{x}_t = \mathbf{v}_t, \qquad \mathbf{v}_t \sim \mathsf{IN}_{10}\left[\mathbf{0},\mathbf{I}\right] \quad \text{for } t = 1,...,T, \qquad (2.24)$$

where ten white-noise processes are generated, denoted $x_{i,t}$, five of which feature in the DGP. Two DGPs are assessed, including a linear DGP in which we set $\beta_{6,0} = \beta_{7,0} = \beta_{8,0} = 0$, and a nonlinear DGP in which $\beta_{i,0} \neq 0$ for $i = 1,...,8$. We can calculate $\boldsymbol{\beta}$ by specifying t-values, due to orthogonality. A t-test of $\mathsf{H}_0 : \beta_k = 0$ is given by:

$$\mathsf{t}_k = \frac{\beta_k}{\sigma_{\beta_k}}. \qquad (2.25)$$

As $\beta_k = \sigma_{x_k y}/\sigma_{x_k}^2$ and $\sigma_{\beta_k}^2 = \sigma_\varepsilon^2/\left(T\sigma_{x_k}^2\right)$, this implies:

$$\mathsf{t}_k = \frac{\beta_k}{\sigma_{\beta_k}} = \frac{\beta_k}{\sqrt{\sigma_\varepsilon^2/\left(T\sigma_{x_k}^2\right)}} = \sqrt{T}\beta_k \frac{\sigma_{x_k}}{\sigma_\varepsilon}. \qquad (2.26)$$

As $\sigma_\varepsilon = 1$ and $\sigma_{x_k} = 1$ for x_k, population t-values of 2, 3, 4, 6 and 8 result in $\beta_{1,0} = 2/\sqrt{T}$, $\beta_{2,0} = 3/\sqrt{T}$, $\beta_{3,0} = 4/\sqrt{T}$, $\beta_{4,0} = 6/\sqrt{T}$, and $\beta_{5,0} = 8/\sqrt{T}$ respectively.

For the nonlinear DGP, to determine $\beta_{6,0}$ we calculate $\sigma_{x_k}^2$ as:

$$\begin{aligned} \sigma_{x_k}^2 &= E\left[\left(x_{1t}^2\right)^2\right] = E\left[x_{1t}^4\right] = 3 \\ \therefore \quad \mathsf{t}_k &= \sqrt{3T}\beta_k \qquad (2.27) \end{aligned}$$

and so we shall set $\beta_{6,0} = 4/\sqrt{3T}$ for a t-value of 4. For $\beta_{7,0}$ and $\beta_{8,0}$:

$$\begin{aligned} \sigma_{x_k x_j} &= E\left[\left(x_{1t}x_{2t}\right)^2\right] = 1 \\ \therefore \quad \mathsf{t}_k &= \sqrt{T}\beta_k \qquad (2.28) \end{aligned}$$

and so we shall set $\beta_7 = 3/\sqrt{T}$ and $\beta_8 = 4/\sqrt{T}$ for t-values of 3 and 4 respectively.

The first experiment that we undertake is the benchmark linear model in which the GUM is an ADL(1,1) model which includes, as non-DGP variables, the lagged endogenous variable, y_{t-1}, the strongly exogenous variables, $x_{6,t}, ..., x_{10,t}$, and the first lags of all regressors:

$$y_t = \pi_{0,1} y_{t-1} + \sum_{k=1}^{10} \sum_{i=0}^{1} \pi_{k,i} x_{k,t-i} + \pi_{0,0} + u_t, \qquad u_t \sim \mathsf{IN}\left[0, \sigma_u^2\right]. \tag{2.29}$$

In (2.29), 17 out of the 22 regressors are nuisance, as $\beta_{6,0} = \beta_{7,0} = \beta_{8,0} = 0$.

The second experiment examines a nonlinear general model based on the level 1 transformations generated by RETINA. (2.30) outlines the GUM: there are a total of 62 regressors, 57 of which are nuisance using the linear DGP.[11]

The final experiment that we undertake assesses the properties of the selection algorithms when the DGP contains the nonlinear terms ($\beta_{i,0} \neq 0$ for $i = 6, 7, 8$) and the GUM is nonlinear, as given in (2.30). In this case, 54 of the 62 regressors are nuisance.

$$\begin{aligned} y_t \;=\; & \pi_{0,1} y_{t-1} + \sum_{k=1}^{10} \sum_{i=0}^{1} \pi_{k,i} x_{k,t-i} + \sum_{k=1}^{5} \gamma_k x_{k,t}^2 + \sum_{k=1}^{5} \delta_k \left(\frac{1}{x_{k,t}}\right) \\ & + \sum_{j=1}^{4} \sum_{l=j+1}^{5} \kappa_{jl} \left(x_{j,t} x_{l,t}\right) + \sum_{j=1}^{4} \sum_{l=j+1}^{5} \mu_{jl} \left(\frac{1}{x_{j,t} x_{l,t}}\right) \\ & + \sum_{j=1}^{4} \sum_{l=j+1}^{5} \phi_{jl} \left(\frac{x_{j,t}}{x_{l,t}}\right) + \pi_{0,0} + \epsilon_t, \qquad \epsilon_t \sim \mathsf{IN}\left[0, \sigma_\epsilon^2\right]. \end{aligned} \tag{2.30}$$

In the Monte Carlo experiments, two sample sizes are examined, $T = 100$ and $T = 1,000$. A further 20 and 200 observations are generated respectively to assess out-of-sample forecast accuracy. $M = 100$ replications are undertaken to reduce the probability of a zero draw.

Results for the orthogonal Monte Carlo experiments

Tables 2.8 and 2.9 record the selection probabilities for a sample size of 100 and 1,000 respectively. The upper half of the tables record overview statistics, including the prob-

[11] In the GUM, the squared-inverses are excluded because the white-noise processes have a zero mean. Also, to narrow down the number of regressors in the GUM, we discard the nonlinear transformations for the strongly exogenous variables $x_{6,t}, ..., x_{10,t}$, only ${}_5C_2$ ratios are included in the GUM as opposed to all permutations, and the first lags of the level 1 transformations are also excluded.

ability that the exact DGP is selected, the probability that an irrelevant variable is included in the selected model, the probability that a relevant variable is excluded from the selected model, the average number of relevant variables retained (denoted non-null rejection), the average number of irrelevant variables retained (denoted null rejection), and finally, two forecast statistics including RMSFE and mean absolute error (MAE). Note that MAPE is not recorded because division by near-zero observations will result in an erratic statistic. The lower half of the tables record the selection probabilities for the non-central t-statistics. As well as recording results for the three experiments outlined above, the retention probabilities for a single t-test on the regressors in the DGP with no selection are reported (denoted simulated: these experiments are undertaken in PcNaive, see Doornik and Hendry, 2001b, at the 5% and 1% significance level, providing the benchmark 'inference costs' against which PcGets and RETINA should be judged).

Figure 2.3 records the probability of retaining the variables in the GUM for the linear model, with the horizontal axis representing the variables, commencing with $1 = y_{t-1}$, $2 = x_t$, $3 = x_{t-1}$, to 22 = intercept. The retention probabilities of all non-DGP variables are fairly constant, at approximately 5% for the liberal strategy and 1% for the conservative strategy. The PcGets selection probabilities are in accordance with the results of Krolzig and Hendry (2001). As RETINA always retains the intercept, retention probabilities for RETINA are calculated excluding the intercept to avoid biasing the null rejection frequency upwards. The retention probabilities for RETINA are lower than the conservative strategy and the probability of retaining relevant variables is also much lower for RETINA, particularly for low non-centralities. With a non-centrality of 8, the retention probabilities are near unity for both programs. RETINA has a much more stringent selection criterion than the PcGets conservative strategy, and so to accurately compare the two algorithms PcGets would ideally be run at RETINA's null rejection rate to see which has a higher power. As RETINA does not have an explicit, constant, null rejection frequency, this is infeasible. While the probability of excluding a relevant variable is much higher for RETINA, the probability of including an irrelevant variable is much lower, demonstrating the trade-off between null and non-null rejection frequencies.

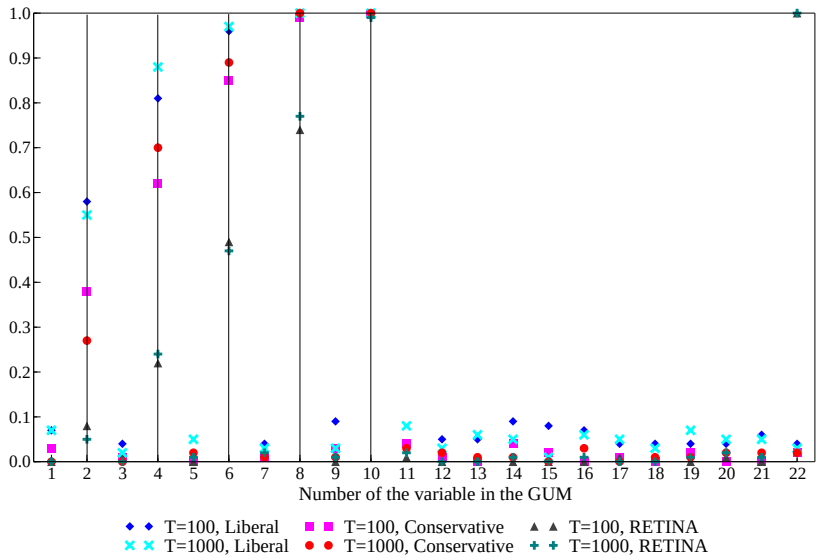

Figure 2.3: The probability of retaining the GUM variable in the specific model, for the linear GUM and linear DGP

Figure 2.4 plots the retention probabilities of the variables in the GUM for the linear DGP and nonlinear GUM.[12] The inclusion of a substantial number of nonlinear and interaction terms does not dramatically alter the properties of either the PcGets or RETINA selection algorithms. The DGP is found less often than the linear case, in accordance with a larger GUM. Also, the probability of not retaining a DGP variable increases, as does the probability of including non-DGP variables, although this effect is much less marked for RETINA. More stringent strategies are preferable for this GUM, because the null rejection frequency becomes more important as the number of variables increase. A sensible way to proceed with PcGets when analysing this type of problem would be to tighten significance levels in direct relation to the number of variables in the GUM. (See Campos *et al.*, 2003, for a discussion on optimal selection strategies.) This suggests that RETINA may be designed for much larger problems than PcGets.

Finally, we assess the case in which nonlinear terms are contained in the DGP. Figure 2.5 plots the retention probabilities, where the nonlinear terms in the DGP are numbered 22 for $x_{1,t}^2$, 36 for $x_{2,t}x_{3,t}$ and 41 for $x_{4,t}x_{5,t}$. The results are similar to those for the linear DGP, although the probability that a DGP variable is not retained is higher for PcGets.

[12]The number ordering on the x-axis proceeds in the order; lagged dependent variable, levels and their lags, squares, inverses, cross-products, cross-inverses, ratios and finally the intercept. The interaction terms are ordered x_1x_2, x_1x_3,..., x_2x_3,... etc. in logical order.

The non-null rejection frequency falls with the larger DGP in all cases, but it is most marked for RETINA where the probabilities of retaining even the largest non-centralities fall. Comparing the results for PcGets to simulations given the known DGP, for smaller t-values there are some small costs of search, but the behaviour of the nonlinear variables is not dramatically different to that of the linear variables. A comparison of RETINA to the simulated DGP is not meaningful as the size of RETINA is not well calibrated. However, assuming a null rejection frequency of 0.001, comparisons with simulated values indicate that again the costs of search are not too high.

Assessing the algorithms on forecast performance is difficult as no clear ranking emerges from RMSFE or MAE. Given that the DGP is stationary, the best forecasting model will be the best in-sample model, which will be the model that most closely captures the DGP. Irrelevant nonlinear functions are likely to be most detrimental to forecasting. We find the linear PcGets models perform very well on RMSFE criteria. RETINA delivers the best forecasts based on MAE for the linear DGP and nonlinear GUM, but the worst forecasts when the DGP is nonlinear, for $T = 100$, indicating that it is difficult to rank the methods on forecast performance.

The Monte Carlo evidence suggests that we can treat orthogonal problems that are nonlinear in the variables (but linear in the parameters) in exactly the same way as linear problems. The size of the tests will need to be controlled if starting with an overly general GUM. If the GUM is mis-specified by a DGP variable not being included in the GUM, both algorithms would face problems. If there is severe mis-specification, PcGets would not commence the path search, whereas RETINA would produce a mis-specified model. If there is parameter change, the problem would be most acute at the forecast origin, in which case both algorithms would suffer. Both algorithms would indicate in-sample parameter non-constancy: PcGets would do so through the reliability weightings and for RETINA, very different local best models would be reported for the six sub-sample combinations.

Table 2.8: Summary statistics and selection probabilities, orthogonal model, $T = 100$

$T = 100$			*Linear GUM,*			*Nonlin GUM,*					*Nonlin GUM,*		
	Simulated		*linear DGP*			*linear DGP*			*Simulated*		*nonlin DGP*		
Strategy	*0.05*	*0.01*	*Lib.*	*Con.*	*Ret.*	*Lib.*	*Con.*	*Ret.*	*0.05*	*0.01*	*Lib.*	*Con.*	*Ret.*
DGP found			0.20	0.15	0.02	0.04	0.03	0.00			0.00	0.00	0.00
Non-DGP inc.			0.58	0.16	0.05	0.81	0.26	0.06			0.84	0.32	0.06
DGP not inc.			0.53	0.77	0.98	0.82	0.95	1.00			0.93	0.98	1.00
Non-null rejection			0.86	0.76	0.52	0.78	0.64	0.39			0.74	0.57	0.27
Null rejection			0.053	0.009	0.007	0.058	0.008	0.001			0.063	0.009	0.001
RMSFE			1.04	1.04	1.13	1.62	1.11	1.19			1.38	1.49	1.37
MAE			2.88	2.87	2.83	3.33	3.09	2.41			2.62	3.50	4.04
t=2	0.47	0.21	0.58	0.38	0.08	0.36	0.18	0.03	0.47	0.24	0.40	0.17	0.03
t=3	0.82	0.63	0.81	0.62	0.22	0.59	0.42	0.16	0.81	0.59	0.63	0.40	0.06
t=4	0.97	0.90	0.96	0.85	0.49	0.75	0.54	0.29	0.96	0.86	0.72	0.51	0.19
t=6	1.00	0.99	1.00	0.99	0.74	0.99	0.97	0.52	1.00	0.99	0.99	0.97	0.51
t=8	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.94	1.00	1.00	1.00	1.00	0.77
t=4 $\left(x_1^2\right)$									0.93	0.83	0.73	0.57	0.20
t=3 (x_2x_3)									0.78	0.57	0.49	0.32	0.08
t=4 (x_4x_5)									0.94	0.84	0.88	0.68	0.12

Table 2.9: Summary statistics and selection probabilities, orthogonal model, $T = 1,000$

$T = 1,000$			*Linear GUM,*			*Nonlin GUM,*					*Nonlin GUM,*		
	Simulated		*linear DGP*			*linear DGP*			*Simulated*		*nonlin DGP*		
Strategy	*0.05*	*0.01*	*Lib.*	*Con.*	*Ret.*	*Lib.*	*Con.*	*Ret.*	*0.05*	*0.01*	*Lib.*	*Con.*	*Ret.*
DGP found			0.21	0.15	0.03	0.06	0.12	0.00			0.05	0.06	0.00
Non-DGP var. inc.			0.54	0.17	0.04	0.83	0.27	0.07			0.83	0.28	0.06
DGP var. not inc.			0.55	0.82	0.97	0.57	0.83	1.00			0.73	0.94	1.00
Non-null rejection			0.88	0.78	0.54	0.88	0.77	0.38			0.87	0.77	0.29
Null rejection			0.045	0.012	0.003	0.048	0.007	0.001			0.041	0.007	0.001
RMSFE			1.12	1.15	1.27	1.54	1.58	1.33			1.42	1.55	1.49
MAE			3.12	3.04	3.29	4.11	4.17	3.89			2.96	3.33	3.69
t=2	0.50	0.27	0.55	0.27	0.05	0.52	0.28	0.04	0.51	0.29	0.49	0.27	0.03
t=3	0.86	0.67	0.85	0.70	0.24	0.85	0.69	0.13	0.85	0.66	0.82	0.71	0.08
t=4	0.99	0.93	0.97	0.89	0.47	0.95	0.87	0.32	0.98	0.91	0.97	0.90	0.24
t=6	1.00	1.00	1.00	1.00	0.77	1.00	1.00	0.56	1.00	1.00	1.00	1.00	0.49
t=8	1.00	1.00	1.00	1.00	0.99	1.00	1.00	0.97	1.00	1.00	1.00	1.00	0.86
t=4 $\left(x_1^2\right)$									0.98	0.92	0.92	0.87	0.25
t=3 (x_2x_3)									0.86	0.67	0.75	0.65	0.11
t=4 (x_4x_5)									0.97	0.91	0.99	0.92	0.19

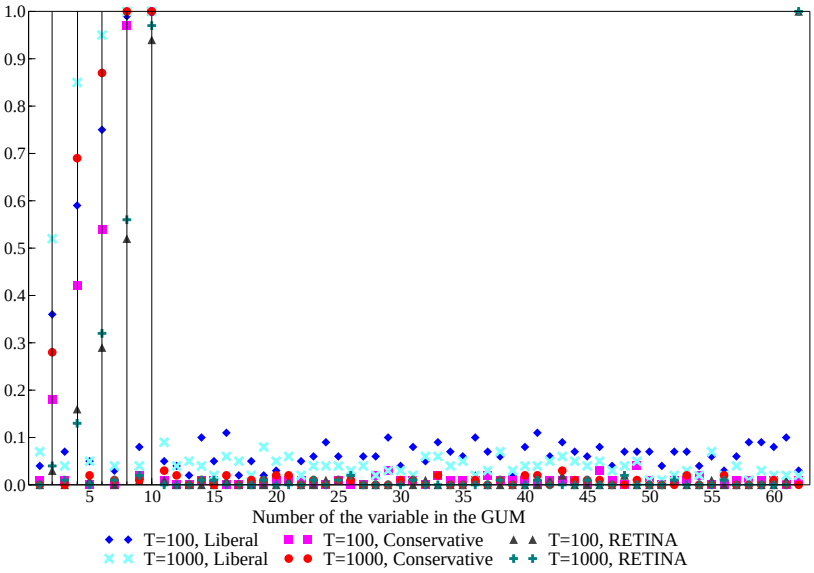

Figure 2.4: The probability of retaining the GUM variable in the specific model, for the nonlinear GUM and linear DGP

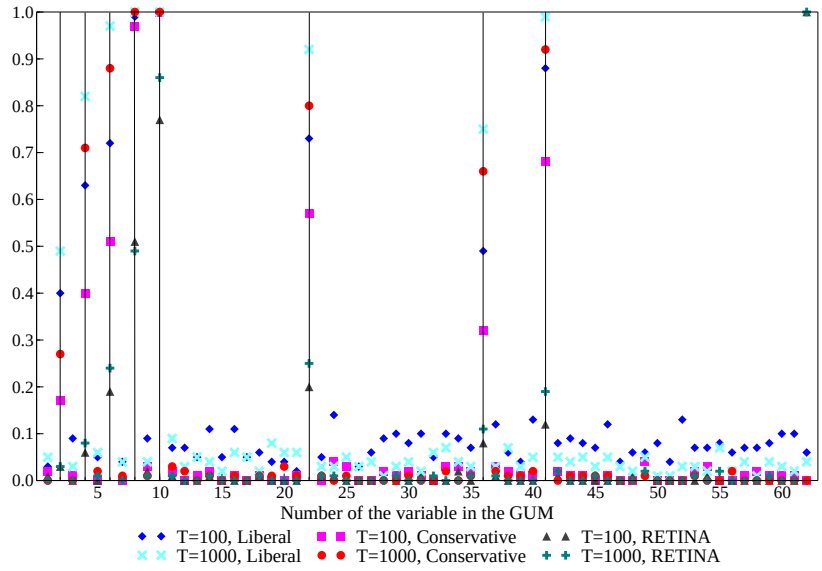

Figure 2.5: The probability of retaining the GUM variable in the specific model, for the nonlinear GUM and nonlinear DGP

2.6.2 Monte Carlo evidence on the non-orthogonal model

Collinearity can be problematic for any selection procedure but is common in economic applications. Hence, we investigate the properties of PcGets and RETINA in a setting in which the regressors are highly collinear. Chapter 3 addresses the problem of collinearity in more detail. The DGP is given by:

$$\begin{aligned} y_t &= \sum_{j=1}^{J} \beta_j x_{j,t} + \epsilon_t, \qquad \epsilon_t \sim \mathsf{IN}\left[0,1\right] \\ \mathbf{x}_t &= \mathbf{10} + \boldsymbol{\nu}_t, \qquad \boldsymbol{\nu}_t \sim \mathsf{IN}_2\left[\mathbf{0},\mathbf{I}\right], \end{aligned} \tag{2.31}$$

for $t = 1, ..., T$, where $J = 1$ or 2. We also consider a nonlinear DGP given by:

$$y_t = \beta_1 x_{1,t}^2 + \beta_2 x_{1,t} x_{2,t} + \epsilon_t. \tag{2.32}$$

The GUM is given by:

$$\begin{aligned} y_t &= \beta_0 + \beta_1 x_{1,t} + \beta_2 x_{2,t} + \beta_3 x_{1,t}^2 + \beta_4 x_{2,t}^2 + \beta_5 \frac{1}{x_{1,t}} + \beta_6 \frac{1}{x_{2,t}} + \beta_7 \frac{1}{x_{1,t}^2} \\ &\quad + \beta_8 \frac{1}{x_{2,t}^2} + \beta_9 x_{1,t} x_{2,t} + \beta_{10} \frac{1}{x_{1,t} x_{2,t}} + \beta_{11} \frac{x_{1,t}}{x_{2,t}} + \beta_{12} \frac{x_{2,t}}{x_{1,t}} + \varepsilon_t, \end{aligned} \tag{2.33}$$

where $\varepsilon_t \sim \mathsf{IN}\left[0, \sigma_\varepsilon^2\right]$. The sample size is $T = 100$ and 1,000 and the number of replications, $M = 1,000$. As the mean of $\mathbf{x}_t$ is non-zero, substantial collinearity is generated between the linear and nonlinear regressors which will require very large parameter values to generate significant non-central t-values. For example, if $x_t = \mu + v_t$, the regression $y_t = \alpha_1 x_t + \alpha_2 x_t^2 + \eta_t$ will result in:

$$\mathsf{E}\left[T^{-1}\left(\mathbf{X}'\mathbf{X}\right)\right] = \begin{bmatrix} \mu^2 + 1 & \mu^3 + 3\mu \\ \mu^3 + 3\mu & 3 + \mu^4 + 6\mu^2 \end{bmatrix}, \tag{2.34}$$

with:

$$\mathsf{E}\left[T^{-1}\left(\mathbf{X}'\mathbf{X}\right)\right]^{-1} = \frac{1}{3 + \mu^4} \begin{bmatrix} 3 + \mu^4 + 6\mu^2 & -\left(\mu^3 + 3\mu\right) \\ -\left(\mu^3 + 3\mu\right) & \mu^2 + 1 \end{bmatrix}. \tag{2.35}$$

With $\mu = 10$, the correlation between x_t and x_t^2 would be 0.99. Given such magnitudes of correlation between the regressors, the coefficient estimates will be strongly driven by the covariance matrix. Observe that the problems of high collinearity can be mitigated by reducing μ or increasing σ_v^2.

Results for the non-orthogonal Monte Carlo experiments

We initially consider the DGP given by $J = 1$ in (2.31) with $\beta_1 = 0.4$ for $T = 100$ and $\beta_1 = 0.12649$ for $T = 1,000$ (corresponding to a non-centrality of 4 in the orthogonal model). Table 2.10 summarises the rejection frequencies for all Monte Carlo experiments undertaken and Table 2.16 in the Appendix records the retention probabilities for the relevant and irrelevant variables for RETINA, PcGets and the simulated DGP from PcNaive for this experiment. The simulated model undertakes individual t-tests on the GUM with no selection, at the 5% and 1% significance levels. The estimated t-values from the simulated model are also recorded, and the high correlation between $x_{1,t}$ and $x_{1,t}^2$ imply that the t-statistic on the relevant variable is insignificant. The retention probabilities of irrelevant variables are too high, with retention rates close to those of the relevant variable for PcGets. The null rejection frequency is approximately 25% for the liberal strategy, 14% for the conservative strategy and 10% for RETINA. The non-null rejection frequency of PcGets is much higher than that of RETINA, which retains $x_{1,t}$ less than 1% of the time, although it picks up $x_{1,t}^2$, which is highly correlated, 35% of the time. RETINA picks up interactions between x_1^2 and x_1^{-2}, which is also highly correlated, although most retention probabilities of non-DGP variables are very small.

To ensure a non-null rejection frequency near unity despite collinearity, very large coefficient values are imposed on the DGP. We can determine values of β_1 that deliver a non-centrality of approximately 4 using a simulated GUM in PcNaive. Coefficients of 1,100 for $T = 100$ and 180 for $T = 1,000$ are required. Table 2.17 in the Appendix records the retention probabilities, with Table 2.10 summarising. RETINA always retains the DGP and rarely retains irrelevant variables, exhibiting good null rejection frequencies given the model design in which a high non-null rejection frequency is ensured. The probability of retaining relevant variables is good for PcGets, but the null rejection frequencies are too high because of high correlations.

A further experiment undertaken examines the case where $J = 2$ in (2.31). Again, large coefficients are required to ensure significant t-statistics for the DGP variables. The experiments are based on non-centralities of approximately 3 and 6 for $x_{1,t}$ and

Table 2.10: Summary statistics: retention probabilities of non-orthogonal experiments

	RETINA		*PcGets Lib.*		*PcGets Cons.*	
Sample	*100*	*1,000*	*100*	*1,000*	*100*	*1,000*
Non-null rejection frequency						
$J=1;\beta_1=0.4$	0.066	0.080	0.331	0.347	0.190	0.190
$J=1;\psi_{\beta_1}\approx 4$	1.000	1.000	0.980	0.993	0.984	0.988
$J=2;\psi_{\beta_1}\approx 3,\psi_{\beta_2}\approx 6$	0.987	0.998	0.970	0.984	0.970	0.978
Non-lin DGP; $\psi_{\beta_i}\approx 4$	0.995	1.000	0.964	0.986	0.969	0.988
$J=1;\psi_{\beta_1}\approx 4$ de-mean	0.812	0.867	0.970	0.972	0.870	0.920
Null rejection frequency						
$J=1;\beta_1=0.4$	0.100	0.096	0.254	0.250	0.142	0.148
$J=1;\psi_{\beta_1}\approx 4$	0.003	0.002	0.153	0.138	0.044	0.033
$J=2;\psi_{\beta_1}\approx 3,\psi_{\beta_2}\approx 6$	0.666	0.592	0.165	0.138	0.068	0.046
Non-lin DGP; $\psi_{\beta_i}\approx 4$	0.047	0.079	0.165	0.143	0.063	0.042
$J=1;\psi_{\beta_1}\approx 4$ de-mean	0.026	0.018	0.052	0.053	0.012	0.012

Notes: $\psi_\beta=\mathsf{E}[|\mathsf{t}_\beta|]$ is the non-centrality of the regressor.

$x_{2,t}$ respectively, which require coefficients of $\beta_1=800$, $\beta_2=1,600$ for $T=100$ and $\beta_1=140$, $\beta_2=280$ for $T=1,000$. Table 2.18 in the Appendix records the retention probabilities for RETINA and PcGets, and Figure 2.6 plots the retention probabilities.[13] The results for PcGets are similar to the single variable case. PcGets does overfit, as the probability of retaining irrelevant variables is too high, but the non-null rejection frequency is almost unity. In contrast to the previous results, the null rejection frequency of RETINA is very high. More variables are retained than would be expected given its emphasis on parsimony. With a large correlation matrix the program struggles to control for the correlation between the variables in order to distinguish the DGP variables, and it retains some irrelevant variables with a probability close to unity.

A fourth experiment undertaken considers the nonlinear DGP outlined in (2.32). To deliver non-centralities of approximately 4, coefficients of $\beta_1=17$ and $\beta_2=40$ for $T=100$, and $\beta_1=2.9$ and $\beta_2=7$ for $T=1,000$, are required. Retention probabilities are reported in Table 2.19 in the Appendix. Results for PcGets are analogous to the linear DGP case, with good non-null rejection frequencies but null rejection frequencies that are too high. For RETINA, the problems observed in the previous experiment are mitigated, with the nonlinear relevant variables being retained with a high probability and a null

[13] The number ordering on the x-axis proceeds in the order: x_1, x_2, x_1^2, x_2^2, $\frac{1}{x_1}$, $\frac{1}{x_2}$, $\frac{1}{x_1^2}$, $\frac{1}{x_2^2}$, x_1x_2, $\frac{1}{x_1x_2}$, $\frac{x_1}{x_2}$, $\frac{x_2}{x_1}$, intercept.

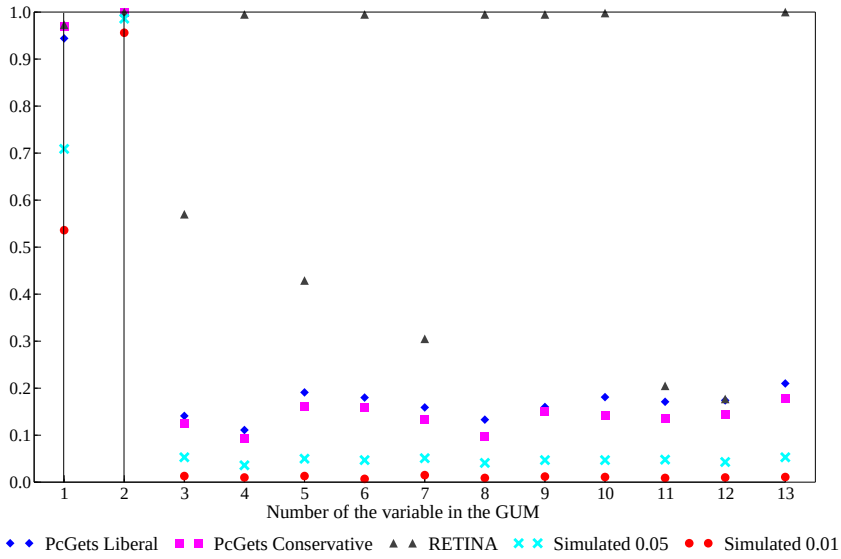

Figure 2.6: The probability of retaining the GUM variable in the specific model; DGP= $\beta_1 x_{1,t} + \beta_2 x_{2,t} + \epsilon_t$, where $\beta_1 = 800$ and $\beta_2 = 1,600$, for $T = 100$.

rejection frequency of approximately 5% for $T = 100$. The null rejection frequency is much larger than in the orthogonal case due to the smaller GUM under consideration. The results suggest that RETINA performs better when the DGP is nonlinear.

The non-orthogonal Monte Carlo evidence highlights the problems with model selection procedures in the presence of nonlinear functions. The substantial collinearity that nonlinear transformations can generate is a problem for both algorithms, resulting in an increase in the null rejection frequency and a reduction in the non-null rejection frequency. While the RETINA selection algorithm should account for collinearity, the evidence in Figure 2.6 indicates poor null rejection frequencies for RETINA when two linear variables enter the DGP. The solution to the problem is to transform the regressors to a near orthogonal representation. This is achieved by 'double de-meaning' the variables, which is discussed in detail in Chapter 3. For $x_{1,t}^2$, the transformation would be $\widetilde{x_{1,t}^2} = (x_{1,t} - \overline{x}_1)^2 - 1$ where $\widetilde{x_{1,t}^2}$ is the de-meaned variable and $\overline{x}_1$ is the mean of $x_{1,t}$. Both the means of $x_{1,t}$ and of $x_{1,t}^2$ need to be removed to eliminate collinearity. Results for both RETINA and PcGets for a range of non-centralities for $J = 1$ in (2.31) are recorded in Table 2.20 in the Appendix. The results indicate that the rejection frequencies of PcGets improve dramatically and, in fact, match previous results for a linear model. Figures 2.7 and 2.8 record the rejection frequencies of the models when all vari-

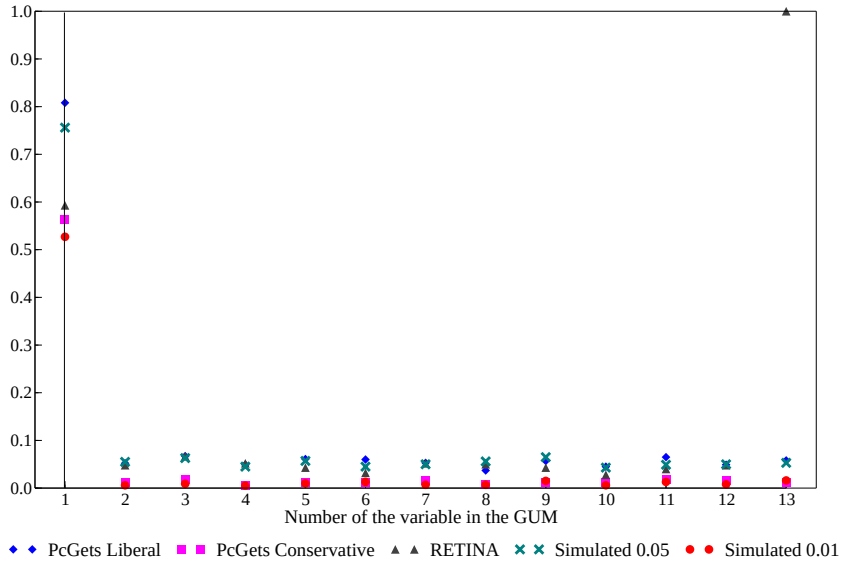

Figure 2.7: The probability of retaining the GUM variable in the specific model for the de-meaned non-orthogonal case; DGP= $\beta_1 x_{1,t} + \epsilon_t$, where $\beta_1 = 0.3$, for $T = 100$.

ables are de-meaned for the 1- and 2- variable DGP respectively, for a non-centrality of 3. The null rejection frequency of PcGets is constant and at theoretical levels. RETINA has a non-constant null rejection frequency across non-centralities that falls as the non-null rejection frequency increases, but the null rejection frequency is much higher than the Monte Carlo experiments from the orthogonal model suggest. This is due to the smaller GUM, as the parsimony benefits of RETINA are achieved in larger GUMs. The non-null rejection frequency of RETINA is lower than the PcGets conservative strategy at non-centralities greater than two.

Collinearity is problematic for both model selection algorithms but a solution is offered: double de-meaning results in a near-orthogonal specification, which implies that nonlinear functions can be analysed in a similar way to linear functions. Both PcGets and RETINA perform well in an orthogonal setting. The null rejection frequency of RETINA depends on the number of variables in the GUM, providing a tighter criterion when the GUM is large. Moreover, RETINA has a lower non-null rejection frequency for smaller non-centralities. PcGets is more flexible in that the null rejection frequency is controlled regardless of the GUM size, and the significance levels can be set by the user depending on the problem under consideration.

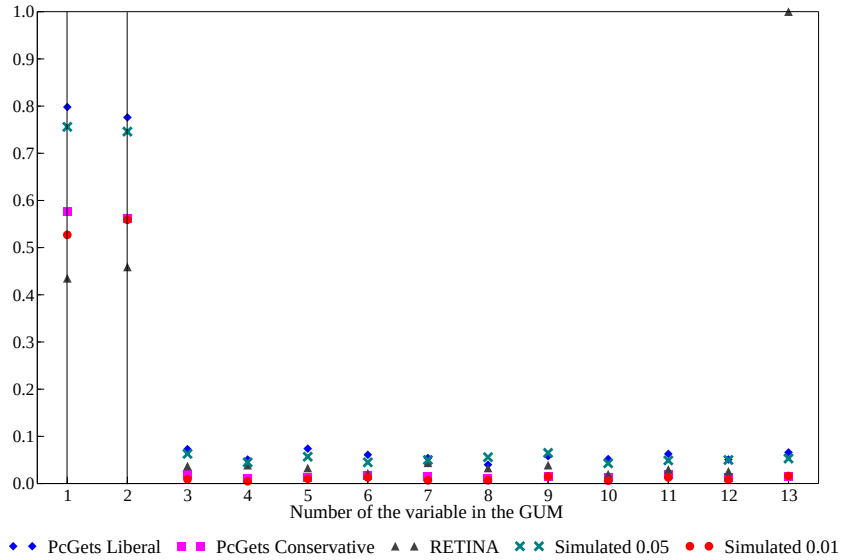

Figure 2.8: The probability of retaining the GUM variable in the specific model for the de-meaned non-orthogonal case; DGP= $\beta_1 x_{1,t} + \beta_2 x_{2,t} + \epsilon_t$, where $\beta_1 = \beta_2 = 0.3$, for $T = 100$.

2.7 Conclusion

The aim of this chapter is to assess two automatic model selection procedures, PcGets and RETINA, and compare the two programs on both cross-section and time-series data in order to establish the advantages of automatically selecting forecasting models. The results are promising; automated methods of model selection have a high level of success and could be dominant in econometric modelling in the future. RETINA is a method of model selection designed primarily to forecast and its predominant feature is the parsimony it can achieve from a highly over-parameterised GUM, whereas PcGets aims to find a congruent, undominated representation of an overly general model. RETINA achieves parsimony via the use of three disjoint sub-samples, but the trade-off is that its selection criteria are very stringent, which can lead to a failure to retain highly significant variables.

Both programs are useful for modelling and forecasting. The ease with which the significance of nonlinearities can be tested with relatively low cost is most advantageous, as demonstrated by the empirical applications. The cross-section results highlight the need to check for outliers. Nonlinear functions may simply be reflecting a few outliers and vastly different results may be obtained by removing these observations. The Monte

Carlo evidence shows that the rejection frequencies of PcGets do not differ substantially because of the inclusion of nonlinear functions of white-noise processes, although one must control the null rejection frequency. The null rejection frequency of RETINA is shown to be much tighter for large GUMs. The evidence for non-orthogonal models is less satisfactory, both for PcGets and RETINA, but a solution is offered. Taking deviations from means mimics a near-orthogonal specification and results in dramatic improvements, both in terms of null and non-null rejection frequencies. This suggests that nonlinearity can be investigated at low cost using automatic model selection algorithms.

2.A Appendix

Table 2.11: Hoover and Perez (1999) data set. Sample: 1959q1-1995q1

Variable	*Times differenced for stationarity*	*Name*
Index of four coincident indicators	1	DCOINC
GNP price deflator	2	GD
Government purchases of goods and services	2	GGEQ
Federal purchases of goods and services	1	GGFEQ
Federal government receipts	2	GGFR
GNP	1	GNPQ
Disposable personal income	1	GYDQ
Gross private domestic investment	1	GPIQ
Total member bank reserves	2	FMRRA
Monetary base (federal reserve bank of St. Louis)	2	FMBASE
M1	1	FM1DQ
M2	1	FM2DQ
Dow Jones stock price	1	FSDJ
Moody's AAA corporate bond yield	1	FYAAAC
Labour force (16 years+, civilian)	1	LHC
Unemployment rate	1	LHUR
Unfilled orders (manufacturing, all industries)	1	MU
New orders (manufacturing, all industries)	2	MO
Personal consumption expenditure	1	GCQ

Table 2.12: Results for Hoover and Perez (1999): Model 1

	Linear			*Nonlinear*		
	PcGets		*RETINA*	*PcGets*		*RETINA*
	Lib.	*Cons.*	*(level 0)*	*Lib.*	*Cons.*	*(level 1)*
RMSFE	175.59	159.27	162.48	148.63	175.59	162.48
MAPE	348.30	249.31	245.01	585.04	348.30	245.011
AIC	9.891	9.945	10.036	9.823	9.891	10.036
$\overline{\mathsf{R}^2}$	0.342	0.299	0.226	0.422	0.342	0.226
Parameters	4	3	2	12	4	2
$\Delta y_{1,t-1}$	−0.448 (0.076)	−0.441 (0.078)	−0.485 (0.081)	−0.417 (0.073)	−0.448 (0.076)	−0.485 (0.081)
intercept			1.078 (13.68)			1.078 (13.68)
$\Delta DCOINC_{t-1}$	9082.82 (2446.7)	7485.39 (2461.4)		9853.0 (2463.5)	9082.8 (2446.7)	
$\Delta GNPQ_{t-1}$	−10243.5 (2335.3)	−8428.0 (2323.6)		−12776 (2294.9)	−10243.5 (2335.3)	
$\Delta FSDJ_t$	572.88 (195.43)			644.65 (228.40)	572.88 (195.43)	
$\Delta^2 MO_{t-1}$				1003.8 (368.48)		
ΔGCQ_t				3980.4 (1699.5)		
$\Delta FYAAAC_t^2$				−9465.7 (3601.4)		
$\frac{1}{\Delta FSDJ}_{t-1}$				−0.384 (0.153)		
$\frac{1}{\Delta^2 MO}_{t}$				−0.036 (0.015)		
$\frac{1}{(\Delta^2 GGEQ)^2}_{t-1}$				−0.00002 (0.00001)		
$\frac{1}{(\Delta FSDJ)^2}_{t-1}$				−0.0001 (0.00005)		
$\frac{1}{(\Delta^2 MO)^2}_{t}$				[t=2.862]		

Notes: The first five rows of the table report summary statistics, with the remaining rows of the table reporting coefficient estimates, with standard errors in parentheses. Square brackets containing t-values are reported where coefficient estimates and standard errors $\leq \left|\times 10^{-6}\right|$.

Table 2.13: Results for Hoover and Perez (1999): Model 2

	Linear			*Nonlinear*		
	PcGets		*RETINA*	*PcGets*		*RETINA*
	Lib.	*Cons.*	*(level 0)*	*Lib.*	*Cons.*	*(level 1)*
RMSFE	207.44	203.12	179.66	221.74	203.12	179.66
MAPE	146.08	134.71	144.81	261.83	134.71	144.81
AIC	10.169	10.283	10.311	10.087	10.283	10.311
$\overline{R^2}$	0.173	0.058	0.039	0.273	0.058	0.039
Parameters	3	1	2	9	1	2
intercept			0.877 (15.69)			0.877 (15.69)
$\Delta^2 GD_t$			−8664.95 (3575.0)			−8664.95 (3575.0)
$\Delta DCOINC_{t-1}$	10102 (2793.3)			9224.5 (2893.5)		
$\Delta GNPQ_{t-1}$	−11422 (2666.1)			−13149 (2695.6)		
$\Delta FSDJ_t$	892.82 (225.39)	626.53 (231.31)		1121.5 (231.45)	626.53 (231.31)	
$\Delta GYDQ_{t-1}$				2988.1 (1653.4)		
$\Delta^2 GGFR^2_{t-1}$				4961.5 (1872.4)		
$\Delta FM1DQ^2_{t-1}$				−88265 (42467)		
$\frac{1}{\Delta MU}{}_t$				−0.316 (0.135)		
$\frac{1}{(\Delta MU)^2}{}_t$				0.0001 (0.00005)		
$\frac{1}{\Delta GGFEQ}{}_{t-1}$				0.00003 (0.00002)		

Notes: The first five rows of the table report summary statistics, with the remaining rows of the table reporting coefficient estimates, with standard errors in parentheses.

Table 2.14: Results for Hoover and Perez (1999): Model 7

	Linear			*Nonlinear*		
	PcGets		*RETINA*	*PcGets*		*RETINA*
	Lib.	*Cons.*	*(level 0)*	*Lib.*	*Cons.*	*(level 1)*
RMSFE	16.711	14.688	14.687	18.800	16.711	14.687
MAPE	114.56	109.37	109.37	91.94	114.56	109.37
AIC	4.990	5.151	5.168	4.895	4.990	5.168
$\overline{R^2}$	0.506	0.405	0.399	0.564	0.505	0.399
Parameters	4	1	2	8	4	2
intercept			−0.003 (1.218)			−0.003 (1.218)
$\Delta FM1DQ_t$	744.095 (92.412)	844.526 (92.452)	844.563 (94.29)	507.729 (127.23)	744.095 (92.411)	844.563 (94.29)
$\Delta DCOINC_{t-1}$	782.670 (208.91)			822.185 (196.60)	782.670 (208.91)	
$\Delta GNPQ_{t-1}$	−866.088 (199.28)			−1022.93 (195.12)	−866.088 (199.28)	
$\Delta FSDJ_t$	70.878 (18.256)			67.928 (17.735)	70.878 (18.256)	
$\Delta^2 MO_{t-1}$				89.901 (30.251)		
$\Delta FM1DQ_{t-1}$				500.539 (136.13)		
$\Delta FM2DQ_t$				255.799 (124.03)		
$\frac{1}{\Delta FSDJ}{}_t$				0.0062 (0.0026)		

Notes: The first five rows of the table report summary statistics, with the remaining rows of the table reporting coefficient estimates, with standard errors in parentheses.

Table 2.15: Results for Hoover and Perez (1999): Model 9

	Linear			*Nonlinear*		
	PcGets		*RETINA*	*PcGets*		*RETINA*
	Lib.	*Cons.*	*(level 0)*	*Lib.*	*Cons.*	*(level 1)*
RMSFE	8.408	7.407	7.403	8.422	7.403	7.403
MAPE	113.56	108.83	108.42	98.136	109.476	108.420
AIC	3.629	3.788	3.805	3.527	3.732	3.805
$\overline{\mathsf{R}^2}$	0.502	0.401	0.396	0.567	0.443	0.396
Parameters	4	1	2	9	3	2
intercept			−0.087 (0.616)			−0.087 (0.616)
$\Delta FM1DQ_t$	373.551 (46.796)	422.876 (46.769)	424.045 (47.69)	315.645 (48.631)	376.600 (50.490)	424.045 (47.69)
$\Delta DCOINC_{t-1}$	389.331 (105.79)			336.817 (103.80)		
$\Delta GNPQ_{t-1}$	−473.694 (100.91)			−487.811 (95.436)	−184.15 (59.210)	
$\Delta FSDJ_t$	35.402 (9.245)			29.173 (9.533)		
$\Delta^2 MO_{t-1}$				42.294 (15.501)		
ΔGCQ_t				149.078 (74.345)	196.372 (72.414)	
$\frac{1}{\Delta FSDJ_t}$				0.0029 (0.0013)		
$\frac{1}{\Delta FM1DQ_{t-1}}$				0.0008 (0.0004)		
$\frac{1}{(\Delta^2 GGEQ)^2_{t-1}}$				[t=-2.188]		

Notes: The first five rows of the table report summary statistics, with the remaining rows of the table reporting coefficient estimates, with standard errors in parentheses. Square brackets containing t-values are reported where coefficient estimates and standard errors $\leq \left|\times 10^{-6}\right|$.

Table 2.16: Retention probabilities for the non-orthogonal model: one variable DGP

	Simulated						*RETINA*		*PcGets*			
Strategy		*0.05*	*0.01*		*0.05*	*0.01*			*Lib.*	*Con.*	*Lib.*	*Con.*
Sample	*t-value*	*100*	*100*	*t-value*	*1,000*	*1,000*	*100*	*1,000*	*100*	*100*	*1,000*	*1,000*
x_1	-0.006	0.049	0.013	0.019	0.036	0.008	0.066	0.080	0.331	0.190	0.347	0.190
x_2	-0.023	0.051	0.007	0.016	0.056	0.006	0.006	0.005	0.251	0.114	0.248	0.125
$(x_1)^2$	0.025	0.053	0.013	0.002	0.048	0.006	0.350	0.393	0.262	0.230	0.280	0.242
$(x_2)^2$	0.011	0.036	0.010	-0.009	0.050	0.009	0.027	0.023	0.191	0.122	0.194	0.130
$\frac{1}{x_1}$	-0.006	0.050	0.013	0.020	0.041	0.008	0.069	0.081	0.248	0.106	0.247	0.105
$\frac{1}{x_2}$	-0.025	0.047	0.007	0.014	0.051	0.008	0.009	0.003	0.265	0.109	0.240	0.119
$\frac{1}{(x_1)^2}$	0.020	0.051	0.015	-0.012	0.039	0.005	0.213	0.232	0.252	0.196	0.260	0.206
$\frac{1}{(x_2)^2}$	0.014	0.041	0.009	-0.008	0.046	0.010	0.034	0.019	0.200	0.123	0.181	0.101
$x_1 x_2$	0.006	0.047	0.012	-0.006	0.046	0.006	0.111	0.084	0.262	0.165	0.279	0.184
$\frac{1}{x_1 x_2}$	0.007	0.047	0.011	-0.003	0.042	0.008	0.113	0.063	0.257	0.118	0.241	0.132
$\frac{x_1}{x_2}$	0.007	0.048	0.009	-0.004	0.044	0.007	0.083	0.069	0.298	0.166	0.274	0.185
$\frac{x_2}{x_1}$	0.006	0.043	0.010	-0.005	0.042	0.007	0.088	0.084	0.268	0.135	0.242	0.126
intercept	0.011	0.053	0.011	-0.023	0.044	0.011	1.000	1.000	0.294	0.119	0.312	0.123

Notes: DGP $= \beta_1 x_{1,t} + \epsilon_t$, where $\beta_1 = 0.4$ for $T = 100$ and $\beta_1 = 0.12649$ for $T = 1,000$.

Table 2.17: Retention probabilities for the non-orthogonal model: one variable DGP

	Simulated						*RETINA*		*PcGets*			
Strategy		*0.05*	*0.01*		*0.05*	*0.01*			*Lib.*	*Con.*	*Lib.*	*Con.*
Sample	*t-value*	*100*	*100*	*t-value*	*1,000*	*1,000*	*100*	*1,000*	*100*	*100*	*1,000*	*1,000*
x_1	3.978	0.898	0.769	3.983	0.957	0.868	1.000	1.000	0.980	0.984	0.993	0.988
x_2	-0.023	0.051	0.007	0.016	0.056	0.006	0.000	0.001	0.168	0.047	0.161	0.034
$(x_1)^2$	0.025	0.053	0.013	0.002	0.048	0.006	0.011	0.009	0.118	0.046	0.114	0.017
$(x_2)^2$	0.011	0.036	0.010	-0.009	0.050	0.009	0.001	0.002	0.121	0.034	0.116	0.023
$\frac{1}{x_1}$	-0.006	0.050	0.013	0.020	0.041	0.008	0.000	0.002	0.166	0.047	0.149	0.041
$\frac{1}{x_2}$	-0.025	0.047	0.007	0.014	0.051	0.008	0.000	0.000	0.169	0.051	0.160	0.035
$\frac{1}{(x_1)^2}$	0.020	0.051	0.015	-0.012	0.039	0.005	0.001	0.000	0.146	0.050	0.119	0.028
$\frac{1}{(x_2)^2}$	0.014	0.041	0.009	-0.008	0.046	0.010	0.000	0.001	0.132	0.028	0.104	0.018
x_1x_2	0.006	0.047	0.012	-0.006	0.046	0.006	0.002	0.005	0.156	0.039	0.138	0.038
$\frac{1}{x_1x_2}$	0.007	0.047	0.011	-0.003	0.042	0.008	0.006	0.001	0.158	0.049	0.135	0.031
$\frac{x_1}{x_2}$	0.007	0.048	0.009	-0.004	0.044	0.007	0.007	0.003	0.163	0.043	0.133	0.036
$\frac{x_2}{x_1}$	0.006	0.043	0.010	-0.005	0.042	0.007	0.005	0.003	0.160	0.044	0.145	0.039
intercept	0.011	0.053	0.011	-0.023	0.044	0.011	1.000	1.000	0.181	0.055	0.179	0.052

Notes: DGP $= \beta_1 x_{1,t} + \epsilon_t$, where β_1 is estimated to give $\psi_{\beta_1} \approx 4$. For $T = 100$, $\beta_1 = 1,100$, and for $T = 1,000$, $\beta_1 = 180$.

Table 2.18: Retention probabilities for the non-orthogonal model: two variable DGP

	Simulated						*RETINA*		*PcGets*			
Strategy		*0.05*	*0.01*		*0.05*	*0.01*			*Lib.*	*Con.*	*Lib.*	*Con.*
Sample	*t-value*	*100*	*100*	*t-value*	*1,000*	*1,000*	*100*	*1,000*	*100*	*100*	*1,000*	*1,000*
x_1	2.891	0.709	0.536	3.102	0.830	0.676	0.973	0.995	0.944	0.944	0.969	0.955
x_2	5.942	0.986	0.956	6.208	0.999	0.995	1.000	1.000	0.996	0.996	0.999	1.000
$(x_1)^2$	0.026	0.053	0.013	0.002	0.048	0.006	0.570	0.403	0.141	0.057	0.124	0.036
$(x_2)^2$	0.011	0.036	0.010	-0.009	0.050	0.009	0.995	1.000	0.111	0.043	0.093	0.017
$\frac{1}{x_1}$	-0.006	0.050	0.013	0.020	0.041	0.008	0.429	0.284	0.191	0.094	0.160	0.068
$\frac{1}{x_2}$	-0.025	0.047	0.007	0.014	0.051	0.008	0.995	1.000	0.180	0.076	0.159	0.051
$\frac{1}{(x_1)^2}$	0.020	0.051	0.015	-0.012	0.039	0.005	0.305	0.144	0.159	0.063	0.133	0.042
$\frac{1}{(x_2)^2}$	0.014	0.041	0.009	-0.008	0.046	0.010	0.995	1.000	0.133	0.041	0.097	0.014
x_1x_2	0.006	0.047	0.012	-0.006	0.046	0.006	0.995	1.000	0.160	0.061	0.150	0.050
$\frac{1}{x_1x_2}$	0.007	0.047	0.011	-0.003	0.042	0.008	0.998	1.000	0.181	0.078	0.142	0.052
$\frac{x_1}{x_2}$	0.007	0.048	0.009	-0.004	0.044	0.007	0.205	0.057	0.171	0.073	0.136	0.052
$\frac{x_2}{x_1}$	0.006	0.043	0.010	-0.005	0.042	0.007	0.177	0.034	0.174	0.067	0.143	0.051
intercept	0.011	0.053	0.011	-0.023	0.044	0.011	1.000	1.000	0.210	0.091	0.177	0.071

Notes: DGP $= \beta_1 x_{1,t} + \beta_2 x_{2,t} + \epsilon_t$, where β_1 and β_2 are estimated to ensure $\psi_\beta \approx 3, 6$. For $T = 100$, $\beta_1 = 800$ and $\beta_2 = 1,600$, and for $T = 1,000$, $\beta_1 = 140$ and $\beta_2 = 280$.

Table 2.19: Retention probabilities for the nonlinear non-orthogonal model

	Simulated						*RETINA*		*PcGets*			
Strategy		*0.05*	*0.01*		*0.05*	*0.01*			*Lib.*	*Con.*	*Lib.*	*Con.*
Sample	*t-value*	*100*	*100*	*t-value*	*1,000*	*1,000*	*100*	*1,000*	*100*	*100*	*1,000*	*1,000*
x_1	-0.036	0.054	0.008	-0.011	0.049	0.007	0.171	0.362	0.179	0.082	0.147	0.051
x_2	-0.043	0.048	0.004	-0.009	0.051	0.007	0.001	0.000	0.162	0.058	0.154	0.041
$(x_1)^2$	4.057	0.836	0.729	4.057	0.943	0.849	0.990	1.000	0.953	0.961	0.986	0.988
$(x_2)^2$	0.034	0.040	0.011	0.008	0.044	0.010	0.003	0.001	0.116	0.035	0.102	0.022
$\frac{1}{x_1}$	-0.036	0.054	0.010	-0.007	0.045	0.008	0.060	0.026	0.188	0.088	0.150	0.049
$\frac{1}{x_2}$	-0.043	0.047	0.004	-0.012	0.054	0.009	0.002	0.000	0.188	0.067	0.160	0.042
$\frac{1}{(x_1)^2}$	0.021	0.049	0.010	-0.012	0.040	0.008	0.039	0.012	0.148	0.069	0.128	0.034
$\frac{1}{(x_2)^2}$	0.030	0.044	0.009	0.007	0.044	0.011	0.006	0.000	0.130	0.035	0.104	0.025
x_1x_2	4.056	0.907	0.824	4.049	0.951	0.876	1.000	1.000	0.975	0.977	0.987	0.989
$\frac{1}{x_1x_2}$	0.031	0.043	0.010	0.011	0.051	0.009	0.186	0.393	0.174	0.064	0.151	0.047
$\frac{x_1}{x_2}$	0.031	0.041	0.011	0.015	0.054	0.012	0.000	0.000	0.171	0.057	0.146	0.043
$\frac{x_2}{x_1}$	0.032	0.044	0.009	0.007	0.052	0.009	0.001	0.000	0.156	0.056	0.153	0.043
intercept	0.043	0.054	0.007	0.011	0.052	0.008	1.000	1.000	0.210	0.083	0.174	0.064

Notes: DGP $= \beta_1 x_{1,t}^2 + \beta_2 x_{1,t} x_{2,t} + \epsilon_t$, where $\beta_1 = 17$ and $\beta_2 = 40$ for $T = 100$, and $\beta_1 = 2.9$ and $\beta_2 = 7$ for $T = 1,000$.

Table 2.20: Retention probabilities for the double de-meaned model

	$\psi = 2$			$\psi = 3$			$\psi = 4$			$\psi = 5$		
$T = 100$	*Lib.*	*Con.*	*RET.*	*Lib.*	*Con.*	*RET.*	*Lib.*	*Con.*	*RET.*	*Lib.*	*Con.*	*RET.*
x_1	0.479	0.234	0.341	0.808	0.564	0.593	0.970	0.870	0.812	0.998	0.981	0.938
x_2	0.053	0.009	0.078	0.052	0.012	0.048	0.053	0.011	0.020	0.053	0.011	0.017
x_1^2	0.062	0.012	0.080	0.067	0.018	0.067	0.060	0.017	0.038	0.060	0.014	0.025
x_2^2	0.048	0.007	0.072	0.049	0.006	0.052	0.050	0.006	0.029	0.049	0.006	0.016
$\frac{1}{x_1}$	0.060	0.009	0.059	0.061	0.011	0.043	0.062	0.012	0.031	0.061	0.013	0.014
$\frac{1}{x_2}$	0.057	0.005	0.052	0.060	0.011	0.032	0.056	0.012	0.024	0.056	0.010	0.014
$\frac{1}{(x_1)^2}$	0.054	0.015	0.072	0.053	0.015	0.052	0.049	0.017	0.037	0.050	0.018	0.016
$\frac{1}{(x_2)^2}$	0.040	0.007	0.069	0.037	0.008	0.050	0.039	0.009	0.021	0.039	0.010	0.014
x_1x_2	0.059	0.011	0.077	0.057	0.011	0.043	0.052	0.009	0.031	0.051	0.009	0.019
$\frac{1}{x_1x_2}$	0.043	0.009	0.048	0.045	0.012	0.028	0.045	0.013	0.014	0.045	0.013	0.009
$\frac{x_1}{x_2}$	0.063	0.017	0.058	0.065	0.019	0.040	0.062	0.018	0.024	0.062	0.018	0.015
$\frac{x_2}{x_1}$	0.047	0.010	0.052	0.049	0.016	0.048	0.047	0.017	0.024	0.047	0.016	0.011
intercept	0.059	0.011	1.000	0.058	0.011	1.000	0.056	0.013	1.000	0.057	0.013	1.000

Notes: DGP $= \beta_1 x_{1,t} + \epsilon_t$. The GUM comprises the double de-meaned regressors, where $\psi = \mathsf{E}[|\mathsf{t}|]$.

Chapter 3

Extending the Boundaries of PcGets: Nonlinear Models

3.1 Introduction

Econometric modelling requires rules for model selection. One systematic approach is the *Gets* methodology automated in PcGets. An extensive assessment of Monte Carlo simulation studies has revealed that the operational characteristics of the PcGets algorithm are excellent across a wide range of states of nature, see Hendry and Krolzig (1999, 2003b, 2005) and Krolzig and Hendry (2001). However, the program is currently designed to select linear models. Economic processes are inherently nonlinear, but we often make a simplifying assumption to reduce the model to a linear representation, leading to potential mis-specification. The LDGP derived in equation (1.5) imposes no restrictions on the functional form of $\mathbf{g}(\mathbf{y}_t)$ and $\mathbf{h}(\mathbf{z}_t)$, and so a model that approximates the LDGP will also need to approximate $\mathbf{g}(.)$ and $\mathbf{h}(.)$.

This chapter develops a strategy for selecting nonlinear models within PcGets, commencing with a new test for nonlinearity and providing solutions to the problems of collinearity generated by nonlinear transformations, extreme observations leading to non-normal distributions, and the use of sufficiently general expansions to approximate the nonlinearity while avoiding excess retention of irrelevant variables. The strategy requires all four developments to be implemented, as exclusion of one aspect of the selection algorithm can be seriously deleterious. The current techniques used to build nonlinear models are mostly specific-to-general, based on a postulated class of models (e.g., the smooth transition regression model (STR), see, *inter alia*, Teräsvirta and Anderson, 1992, Teräsvirta, 1995, and Franses and Teräsvirta, 2001), and the main contribution of this chapter is to provide an alternative methodology for selecting nonlinear models, motivated by the theory of reduction and hence, embedded within the *Gets* framework.

The proposed strategy for selecting nonlinear models commences by specifying a linear GUM that includes all potentially relevant variables. Nonlinearity is then tested for within the batch of diagnostic tests used to ensure congruency. The test procedure is close to the test for heteroskedasticity proposed by White (1980), but by using weighted combinations of the squares and cross-products based on the eigenvectors of the variance-covariance matrix, the test circumvents problems of high dimensional systems and collinearity. Rejection of the null then requires that the process is modelled within the nonlinear framework. As the number of possible nonlinear functions is infinite, an approximating class is required. Given that many nonlinear models, including smooth transition regressions, regime-switching models, neural networks and nonlinear VARs can be approximated by Taylor expansions, polynomials form an appropriate approximating class. The PcGets selection algorithm is then applied to the re-specified nonlinear GUM to select an undominated, congruent, specific model.

There are problems with selecting under nonlinearity that need to be addressed before an operational algorithm can be implemented. First, undertaking a nonlinear transformation of a variable can generate substantial collinearity between the original linear function and the transformed nonlinear function, as demonstrated in Chapter 2. Collinearity is problematic for any selection procedure because the information content of an extra highly collinear variable is very small, yet it disrupts existing information attribution. Any model selection algorithm will struggle to determine the relevant variables, and will therefore select poorly between the relevant and irrelevant variables depending on sampling error. Orthogonality is highly beneficial for model selection in general, both for that reason, and because deleting small, insignificant coefficients leaves the retained estimates almost unaltered. Operational rules are proposed to transform the nonlinear functions to a more orthogonal form prior to undertaking model selection, and this results in greatly improved properties of the selection procedure. Secondly, some classes of nonlinear functions can generate extreme observations and the resulting fat-tailed distributions are problematic because the assumption of normality is often inbuilt for the critical values used by model selection procedures. Nonlinear functions can 'line up' with outliers,

causing those functions to be retained too often. We propose a solution of removing the extreme observations using indicator saturation techniques, developed by Hendry *et al.* (2004), to ensure near normality for inference. This innovative technique also avoids the problem of otherwise undetectable outliers. Finally, the ability of PcGets to handle more variables than observations enables sufficiently general expansions to approximate the nonlinearity. However, we seek to control the problem of excess retention of nonlinear functions due to an over-parameterised GUM by proposing a 'super-conservative' strategy for selecting nonlinear functions.

The structure of the chapter is as follows. Section 3.2 outlines the nonlinearity test and establishes the power of the test under alternative LDGP specifications. Section 3.3 considers the polynomial class of functions and investigates their ability to approximate a STR model. The LSTR model nests the two regime-switching model and captures characteristics of ANNs, Markov-switching models and nonlinear VARs, hence providing a general benchmark model on which to assess the polynomial approximation. Section 3.4 addresses the intrinsic problems of selecting models that are nonlinear in the regressors, including collinearity between linear and nonlinear functions; non-normality whereby a Monte Carlo study highlights the problem of extreme observations for model selection; and the use of a super-conservative strategy to ensure nonlinear functions are retained only when there is definite evidence of nonlinearity in the data. Finally, section 3.5 concludes.

3.2 Functional form testing

An essential component of congruency concerns the validity of the functional form, and therefore a test for nonlinearity is required that can evaluate the 'goodness' of the postulated linear model. The modelling strategy commences by specifying a GUM in which all potentially relevant variables are included in a linear functional form. A test for nonlinearity is then undertaken within the batch of diagnostic tests. If the null of linearity is accepted, the PcGets selection algorithm is applied. If the null is rejected, the GUM is reformulated to include a set of polynomial approximations, and selection is then applied to the more general nonlinear model. Two types of linearity test exist: those that test

against a specific alternative and those that test against any departure from linearity, and, in keeping with the *Gets* methodology, we seek a test against any alternative.

3.2.1 A quadratic approximation test

Consider a general relation:

$$y_t = f(x_{1,t} \dots x_{n,t}) + \epsilon_t, \tag{3.1}$$

where the linear approximation is:

$$y_t = \beta_0 + \beta_1 x_{1,t} + \cdots + \beta_n x_{n,t} + e_t = \beta_0 + \boldsymbol{\beta}' \mathbf{x}_t + e_t. \tag{3.2}$$

To evaluate the goodness of (3.2) requires testing the validity of the functional form approximation. Using the mean value theorem around a fixed point $\mathbf{x}_0$:

$$\begin{aligned} f(\mathbf{x}_t) &= f(\mathbf{x}_t)_{\rfloor \mathbf{x}_t = \mathbf{x}_0} + \frac{\partial f(\mathbf{x}_t)}{\partial \mathbf{x}_t'}_{\rfloor \mathbf{x}_t = \mathbf{x}_0} (\mathbf{x}_t - \mathbf{x}_0) + (\mathbf{x}_t - \mathbf{x}_0)' \frac{\partial^2 f(\mathbf{x}_t)}{\partial \mathbf{x}_t \partial \mathbf{x}_t'}_{\rfloor \mathbf{x}_t = \mathbf{x}^*} (\mathbf{x}_t - \mathbf{x}_0) \\ &= f(\mathbf{x}_0) + \boldsymbol{\beta}' (\mathbf{x}_t - \mathbf{x}_0) + (\mathbf{x}_t - \mathbf{x}_0)' \mathbf{A}^* (\mathbf{x}_t - \mathbf{x}_0) \\ &= \beta_0 + \boldsymbol{\beta}' \mathbf{x}_t + \mathbf{x}_t' \mathbf{A}^* \mathbf{x}_t \end{aligned} \tag{3.3}$$

where $\mathbf{x}^* \in [\mathbf{x}_t, \mathbf{x}_0]$. Thus, a 'natural' test is to consider the importance of the quadratic term $\mathbf{x}_t' \mathbf{A}^* \mathbf{x}_t$ added to (3.2). Since:

$$\mathbf{x}_t' \mathbf{A}^* \mathbf{x}_t = \mathsf{tr}\left(\mathbf{x}_t' \mathbf{A}^* \mathbf{x}_t\right) = \mathsf{tr}\left(\mathbf{A}^* \mathbf{x}_t \mathbf{x}_t'\right), \tag{3.4}$$

for fixed regressors, an exact test would be an F-test of $\boldsymbol{\delta}_1 = \mathbf{0}$ in:

$$y_t = \beta_0 + \boldsymbol{\beta}' \mathbf{x}_t + \boldsymbol{\delta}_1' \mathbf{w}_t + e_t \tag{3.5}$$

where:

$$\mathbf{w}_t = \left(\mathbf{x}_t \mathbf{x}_t'\right)^{v_e} \tag{3.6}$$

and v_e vectorises and selects non-redundant elements from the lower triangle (including the diagonal) of the outer product. This procedure is close to the test for heteroskedasticity proposed by White (1980), where his heteroskedastic-consistent covariance matrix estimator will differ from the conventional formula when the squares and cross-products

of the regressors would be significant if added to the model.

There are three main drawbacks of such a test: first, its high dimensionality; secondly the potentially high collinearity between the elements of $\mathbf{w}_t$; and third, the possibility that the second derivative is not the source of the important departure from linearity, which may depend on asymmetry or skewness and be better reflected in the third derivative. To rectify these potential drawbacks we develop an alternative test.

3.2.2 Improved testing

First, consider the optimal test of $\boldsymbol{\delta}_1 = \mathbf{0}$ when $f(\mathbf{x}_t)$ is the exact quadratic:

$$f(\mathbf{x}_t) = \beta_0 + \boldsymbol{\beta}'\mathbf{x}_t + \mathbf{x}_t'\mathbf{A}\mathbf{x}_t \tag{3.7}$$

and $\mathbf{A}$ is known and symmetric. Let $\mathbf{A} = \mathbf{K}\boldsymbol{\Upsilon}\mathbf{K}'$, where $\boldsymbol{\Upsilon}$ is the matrix of eigenvalues of $\mathbf{A}$ and $\mathbf{g}_t = \mathbf{K}'\mathbf{x}_t$, so that

$$\mathbf{x}_t'\mathbf{A}\mathbf{x}_t = \mathbf{x}_t'\mathbf{K}\boldsymbol{\Upsilon}\mathbf{K}'\mathbf{x}_t = \mathbf{g}_t'\boldsymbol{\Upsilon}\mathbf{g}_t = \sum_{i=1}^{n} \tau_i g_{i,t}^2 = \boldsymbol{\tau}'\mathbf{r}_t \tag{3.8}$$

where $\boldsymbol{\tau}$ is the vector of diagonal elements of $\boldsymbol{\Upsilon}$ and $\mathbf{r}_t$ is the $n \times 1$ vector with elements $g_{i,t}^2$. Then:

$$y_t = \beta_0 + \boldsymbol{\beta}'\mathbf{x}_t + \boldsymbol{\tau}'\mathbf{r}_t + \epsilon_t = \beta_0 + \boldsymbol{\beta}'\mathbf{x}_t + \delta w_t + \epsilon_t \tag{3.9}$$

say, so a t-test of H_0: $\delta = 0$ will be the most powerful test for nonlinearity when $w_t = \boldsymbol{\tau}'\mathbf{r}_t$.

To provide an operational counterpart when $\mathbf{A}$ is unknown, let:

$$\mathbf{x}_t \sim \mathbf{D}_n[\boldsymbol{\mu}, \boldsymbol{\Omega}] \tag{3.10}$$

such that $\boldsymbol{\Omega} = \mathbf{H}\boldsymbol{\Lambda}\mathbf{H}'$ and $\mathbf{z}_t^* = \mathbf{H}'\mathbf{x}_t$ where $\mathbf{H}'\mathbf{H} = \mathbf{I}_n$ so that:

$$\mathbf{z}_t^* \sim \mathbf{D}_n\left[\mathbf{H}'\boldsymbol{\mu}, \boldsymbol{\Lambda}\right]. \tag{3.11}$$

Finally, take deviations from their means, then scale the $z_{i,t}^d = z_{i,t}^* - \overline{z_i^*}$ by the square roots of their corresponding λ_i:

$$\frac{z_{i,t}^d}{\sqrt{\lambda_i}} = z_{i,t}, \tag{3.12}$$

so:

$$\mathbf{z}_t \underset{app}{\sim} \mathbf{D}_n \left[\mathbf{0}, \mathbf{I}\right]. \tag{3.13}$$

Thus, the $z_{i,t}$ are standardised and mutually orthogonal combinations of squares and cross-products of the original $x_{i,t}$.

As (3.4) is a scalar:

$$\mathsf{tr}\left(\mathbf{x}_t'\mathbf{A}^*\mathbf{x}_t\right) = \mathsf{tr}\left(\mathbf{x}_t'\left(\mathbf{A}^*\right)'\mathbf{x}_t\right) = \frac{1}{2}\mathsf{tr}\left(\mathbf{x}_t'\left[\mathbf{A}^* + \left(\mathbf{A}^*\right)'\right]\mathbf{x}_t\right) = \mathsf{tr}\left(\mathbf{x}_t'\mathbf{B}\mathbf{x}_t\right) \tag{3.14}$$

where $\mathbf{B}$ is symmetric. Hence, we can take $\mathbf{A}^*$ to be symmetric without loss of generality. Next:

$$\mathsf{tr}\left(\mathbf{x}_t'\mathbf{A}^*\mathbf{x}_t\right) = \mathsf{tr}\left(\mathbf{x}_t'\mathbf{H}\left(\mathbf{H}'\mathbf{A}^*\mathbf{H}\right)\mathbf{H}'\mathbf{x}_t\right) = \mathsf{tr}\left(\left(\mathbf{H}'\mathbf{A}^*\mathbf{H}\right)\mathbf{z}_t\mathbf{z}_t'\right) = \mathsf{tr}\left(\mathbf{C}\mathbf{z}_t\mathbf{z}_t'\right). \tag{3.15}$$

Under the null of a linear function, $\mathbf{C} = \mathbf{0}$, so for a local alternative, exploiting symmetry, we expand $\mathbf{C}$ around $\mathbf{0}$ as $\mathbf{C} = \boldsymbol{\Delta}_n$ for a diagonal $\boldsymbol{\Delta}_n$ ($= \mathbf{0}$ under the null). Hence, from (3.15):

$$\mathsf{tr}\left(\mathbf{C}\mathbf{z}_t\mathbf{z}_t'\right) \simeq \mathsf{tr}\left(\boldsymbol{\Delta}_n\mathbf{z}_t\mathbf{z}_t'\right) = \mathbf{z}_t'\boldsymbol{\Delta}_n\mathbf{z}_t = \sum_{i=1}^{n}\delta_{1,i}z_{i,t}^2. \tag{3.16}$$

Thus, the test has the same form as that in (3.9), but where $u_{i,t} = z_{i,t}^2$ in place of $w_{i,t}$. Relative to (3.5), there are only n elements in $\mathbf{u}_t$, as opposed to $n\left(n+1\right)/2$, but every element potentially depends on squares and cross-products of every $x_{i,t}$. Thus, the first and second objectives of effecting a major dimensionality reduction and formulating a test in terms of non-collinear variables are achieved. Additional terms from the next sub and super diagonals of $\mathbf{C}$ could also be used; or going in the opposite direction, a scalar test corresponding to (3.9) could be constructed using $\sum_{i=1}^{n} z_{i,t}^2$ as a single regressor.

Under the null, for fixed regressors and $e_t \sim \mathsf{IN}\left(0, \sigma_e^2\right)$, the test of $\boldsymbol{\delta}_1 = \mathbf{0}$ in:

$$y_t = \beta_0 + \boldsymbol{\beta}'\mathbf{x}_t + \boldsymbol{\delta}_1'\mathbf{u}_t + e_t \tag{3.17}$$

where $u_{i,t} = z_{i,t}^2$, is an exact F-test with n degrees of freedom. Under the alternative, the test should have power against quadratic departures that are not orthogonal to the $w_{i,t}$. Finally, it is easy to accommodate the third drawback, and generalise the test for higher-derivative departures from the null by also including $\left\{z_{i,t}^3\right\}$ terms, which leads

again to an exact F-test under the null. When the additional terms $\sum_{j=1}^{n} \delta_{2,j} z_{j,t}^3$ are included, the test is an F-test of $\boldsymbol{\delta}_1 = \boldsymbol{\delta}_2 = \mathbf{0}$, with $2n$ degrees of freedom.

3.2.3 Non-centrality

The easiest case to consider is (3.7), as a t-test of H_0: $\delta = 0$ will be the most powerful test for non-linearity in (3.9):

$$y_t = \beta_0 + \boldsymbol{\gamma}' \mathbf{x}_t + \delta \left(\boldsymbol{\tau}' \mathbf{r} \right)_t + \epsilon_t = \beta_0 + \boldsymbol{\gamma}' \mathbf{x}_t + \delta w_t + \epsilon_t. \tag{3.18}$$

Then letting $\beta_0 = 0$ and $\boldsymbol{\mu} = \mathbf{0}$ for simplicity, so all linear terms have means of zero, and:

$$\mathbf{Q} = \mathbf{I}_T - \mathbf{X} \left(\mathbf{X}'\mathbf{X} \right)^{-1} \mathbf{X}'$$

where:

$$\mathbf{y}' = (y_1 \dots y_T) \quad \text{and} \quad \mathbf{X}' = (\mathbf{x}_1 \dots \mathbf{x}_T),$$

then:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\gamma} + \mathbf{w}\delta + \boldsymbol{\epsilon}, \tag{3.19}$$

so for $\boldsymbol{\epsilon} \sim \mathsf{IN}\left[\mathbf{0}, \sigma_\epsilon^2 \, \mathbf{I}_T\right]$:

$$\sqrt{T}\left(\widehat{\delta} - \delta \right) = \left(T^{-1} \mathbf{w}'\mathbf{Q}\mathbf{w} \right)^{-1} \frac{\mathbf{w}'\mathbf{Q}\boldsymbol{\epsilon}}{\sqrt{T}} \xrightarrow{D} \mathsf{N}\left[0, \sigma_\delta^2\right], \tag{3.20}$$

where:

$$\sigma_\delta^2 = \sigma_\epsilon^2 \text{plim}_{T\to\infty} \left(T^{-1} \mathbf{w}'\mathbf{Q}\mathbf{w} \right)^{-1}. \tag{3.21}$$

After double de-meaning to ensure the squares and cross-products are not highly collinear with the $x_{i,t}$, $\mathbf{Q}\mathbf{w}$ should be approximately $\mathbf{w}$, leading to:

$$\sigma_\delta^2 = \sigma_\epsilon^2 \text{plim}_{T\to\infty} \left(T^{-1} \mathbf{w}'\mathbf{w} \right)^{-1}. \tag{3.22}$$

From (3.8):

$$T^{-1}\mathbf{w}'\mathbf{w} = T^{-1} \sum_{t=1}^{T} \left(\sum_{i=1}^{n} \sum_{j=1}^{n} \tau_i \tau_j r_{i,t}^2 r_{j,t}^2 \right) \to \boldsymbol{\tau}' \mathbf{R} \boldsymbol{\tau} \tag{3.23}$$

and $\mathbf{R}$ is the limit of the matrix with elements $\left\{ T^{-1} \sum_{t=1}^{T} r_{i,t}^2 r_{j,t}^2 \right\}$. Then:

$$\sqrt{T}\left(\widehat{\delta} - \delta \right) \xrightarrow{D} \mathsf{N}\left[0, \sigma_\epsilon^2 \left(\boldsymbol{\tau}' \mathbf{R} \boldsymbol{\tau} \right)^{-1} \right]. \tag{3.24}$$

Under the sequence of local alternatives that $\delta = c/\sqrt{T}$:

$$\mathsf{t}^2_{\delta=0} = \frac{T\widehat{\delta^2}\left(\boldsymbol{\tau}'\mathbf{R}\boldsymbol{\tau}\right)}{\sigma^2_\epsilon} \tag{3.25}$$

which has a local non-centrality parameter:

$$\psi = \frac{c^2\left(\boldsymbol{\tau}'\mathbf{R}\boldsymbol{\tau}\right)}{\sigma^2_\epsilon}. \tag{3.26}$$

To relate the operational test in (3.17) to the optimal, and derive the non-centrality corresponding to (3.26), again for $\beta_0 = 0$, consider:

$$\mathbf{x}'_t\mathbf{A}\mathbf{x}_t = \mathbf{x}'_t\mathbf{H}\left(\mathbf{H}'\mathbf{A}\mathbf{H}\right)\mathbf{H}'\mathbf{x}_t = \mathbf{z}'_t\left(\mathbf{H}'\mathbf{K}\boldsymbol{\Upsilon}\mathbf{K}'\mathbf{H}\right)\mathbf{z}_t = \mathbf{z}'_t\mathbf{C}\mathbf{z}_t, \tag{3.27}$$

so that, using the symmetry of $\mathbf{A}$:

$$\mathbf{C} = \mathbf{H}'\mathbf{A}\mathbf{H} = \boldsymbol{\Delta}_n + \mathbf{D}, \tag{3.28}$$

where $\mathbf{D}$ has a zero diagonal:

$$y_t = \boldsymbol{\beta}'\mathbf{x}_t + \mathbf{z}'_t\boldsymbol{\Delta}_n\mathbf{z}_t + \mathbf{z}'_t\left(\mathbf{C} - \boldsymbol{\Delta}_n\right)\mathbf{z}_t + \epsilon_t = \boldsymbol{\beta}'\mathbf{x}_t + \boldsymbol{\delta}'_1\mathbf{u}_t + \mathbf{z}'_t\mathbf{D}\mathbf{z}_t + \epsilon_t. \tag{3.29}$$

Thus, the closer $\mathbf{D}$ is to $\mathbf{0}$, the less the power loss. If, for example, $\mathbf{A}$ was positive definite, then $\mathbf{H}$ could jointly diagonalise $\mathbf{A}$ and $\boldsymbol{\Omega}$, such that $\boldsymbol{\Delta}_n = \mathbf{I}_n$ and so $\mathbf{D} = \mathbf{0}$ (see, e.g., Hendry, 1995, p.631).

3.2.4 Test power

To compute the power function, we approximate F^k_{T-k} by a chi-squared with n degrees of freedom:

$$\mathsf{F}^k_{T-k}\left(\varphi_{r,\alpha}\right) \sim \chi^2_k\left(\varphi^2_{r,\alpha}\right). \tag{3.30}$$

Next, we relate that non-central χ^2 distribution to a central χ^2 using (see, e.g., Hendry, 1995, p.475):

$$\chi^2_k\left(\varphi^2_{r,\alpha}\right) = h\chi^2_m\left(0\right), \tag{3.31}$$

such that:

$$h = \frac{k + 2\varphi^2_{r,\alpha}}{k + \varphi^2_{r,\alpha}} \quad \text{and} \quad m = \frac{k + \varphi^2_{r,\alpha}}{h}. \tag{3.32}$$

Finally, we calculate the power function of the $\chi^2_k\left(\varphi^2_{r,\alpha}\right)$ test in (3.30) using:

$$\mathsf{P}\left[\chi^2_k\left(\varphi^2_{r,\alpha}\right) > c_\alpha \mid \mathsf{H}_1\right] \simeq \mathsf{P}\left[\chi^2_m\left(0\right) > h^{-1}c_\alpha\right] \tag{3.33}$$

For example, when $k = 20$ and $\varphi^2_{r,\alpha} = 5$, then $h = 70/45 \simeq 1.56$ and $m \simeq 29$ with $c_\alpha \simeq 31.4$ for $\chi^2_{20}\left(0\right)$ so $\mathsf{P}\left[\chi^2_{29}\left(0\right) > 20.1\right] \simeq 0.89$, delivering a reasonable power. Reducing the degrees of freedom of the test would increase power at the same $\varphi_{r,\alpha}$, or more generally there is a trade-off between fewer terms and the value of $\varphi_{r,\alpha}$.

3.2.5 Powerless case

The test will have no power if the departure from linearity is in the direction of $\mathbf{u}_{t,\perp}$, which would require $\mathbf{\Delta}_n = \mathbf{0}_n$ when $\mathbf{D} \neq \mathbf{0}$. This may occur if the $x_{i,t}$s were perfectly orthogonal, such that $\mathbf{\Omega} = \sigma^2_x\mathbf{I}_n$, and the nonlinearity entered in the form of a cross-product. In this case, the matrix of eigenvectors would be the identity matrix, and the resulting linear combinations would exclude cross-product terms. However, the economic relevance of a case in which two variables are perfectly orthogonal but that their cross-product is relevant, is likely to be limited. Furthermore, the test would have no power if the second derivative of $f\left(\cdot\right)$ was zero, but the third was non-zero: this is precisely the reason for seeking to include the additional terms $\sum_{j=1}^{n}\delta_{2,j}z^3_{j,t}$. If the first non-zero derivative is the fourth, the test will have power, as the second derivative will almost certainly be correlated with the fourth. Monte Carlo experiments are undertaken to confirm these conjectures.

3.2.6 Optimal test

We first examine the power of the 'optimal' test, defined as the F-test of the nonlinear regressors in the DGP, which are assumed to be known. This provides the benchmark against which to assess both White's test and the test outlined above, denoted the 'index' test. Consider a polynomial DGP given by:

$$y_t = \beta x^i_t + \epsilon_t, \tag{3.34}$$

where $\epsilon_t \sim \mathsf{IN}[0,1]$ and $x_t \sim \mathsf{IN}[0,1]$, for $t = 1, ..., T$, and $i = 1, ..., 4$. Hence, the four cases include a linear, quadratic, cubic and quartic function. To calculate the non-centrality, since:[1]

$$\frac{\sqrt{T\mathsf{E}\left[\left(x_t^i\right)^2\right]}\left(\widehat{\beta}-\beta\right)}{\sqrt{\sigma_\epsilon^2}} \sim \mathsf{N}[0,1] \tag{3.35}$$

then:

$$\frac{T\left(\widehat{\beta}-\beta\right)^2 \mathsf{E}\left[\left(x_t^i\right)^2\right]}{\sigma_\epsilon^2} \sim \chi_1^2 \tag{3.36}$$

so:

$$\frac{T\widehat{\beta}^2\mathsf{E}\left[\left(x_t^i\right)^2\right]}{\sigma_\epsilon^2} \sim \chi_1^2\left(\varphi_{\alpha,i}^2\right) \tag{3.37}$$

and

$$\varphi_{\alpha,i}^2 = \frac{T\beta^2\mathsf{E}\left[\left(x_t^i\right)^2\right]}{\sigma_\epsilon^2}, \tag{3.38}$$

where $\mathsf{E}\left[\left(x_t\right)^2\right] = 1$; $\mathsf{E}\left[(x_t^2)^2\right] = 3$; $\mathsf{E}\left[(x_t^3)^2\right] = 15$; and $\mathsf{E}\left[(x_t^4)^2\right] = 105$. Using (3.38), values of β can be backed out for a given $\varphi_{\alpha,i}$. We undertake an experiment in which the Monte Carlo non-centralities are calculated for a given β corresponding to the analytic $\varphi_{\alpha,i}$, ranging from $\varphi_{\alpha,i} = 1, ..., 6$, $\forall i = 1, ..., 4$. For example, for $\varphi_{\alpha,i} = 3$ and $T = 100$, β is given by $T^{-\frac{1}{2}}\varphi_{\alpha,1} = 0.3$ for the linear DGP; $(3T)^{-\frac{1}{2}}\varphi_{\alpha,2} = 0.1732$ for the quadratic DGP; $(15T)^{-\frac{1}{2}}\varphi_{\alpha,3} = 0.0775$ for the cubic DGP; and $(105T)^{-\frac{1}{2}}\varphi_{\alpha,4} = 0.0293$ for the quartic DGP. $10,000$ replications of (3.34) are undertaken and the corresponding estimated non-centralities are recorded in Table 3.1.

For the linear DGP, the Monte Carlo non-centralities match the analytic non-centralities. However, as the polynomial order increases, the approximation becomes poorer, with deviations of 10% for the quartic function over intermediate non-centralities. This is a small sample problem; as T increases, the gap between the analytic and Monte Carlo non-centralities narrows. The poor approximation is driven by the skewness and kurtosis

[1] As $\widehat{\beta} = \beta + \left(\sum\left(x_t^i\right)^2\right)^{-1}\sum x_t^i\epsilon_t$, and $\widehat{\beta}^2 = \beta^2 + 2\beta\left(\sum\left(x_t^i\right)^2\right)^{-1}\sum x_t^i\epsilon_t + \left[\left(\sum\left(x_t^i\right)^2\right)^{-1}\sum x_t^i\epsilon_t\right]^2$,

$$\mathsf{E}\left[\widehat{\beta}^2\right] = \beta^2 + \left(\sum\left(x_t^i\right)^2\right)^{-2}\mathsf{E}\left[\sum\left(x_t^i\right)^2\epsilon_t^2\right] = \beta^2 + \left(\sum\left(x_t^i\right)^2\right) \simeq \beta^2 + T\sigma_\epsilon^2\mathsf{E}\left[\left(x_t^i\right)^2\right],$$

where $\widehat{\sigma}_\beta^2 = \widehat{\sigma}_\epsilon^2\left(\sum\left(x_t^i\right)^2\right)^{-1} \simeq \sigma_\epsilon^2\left(\sum\left(x_t^i\right)^2\right)^{-1}$.

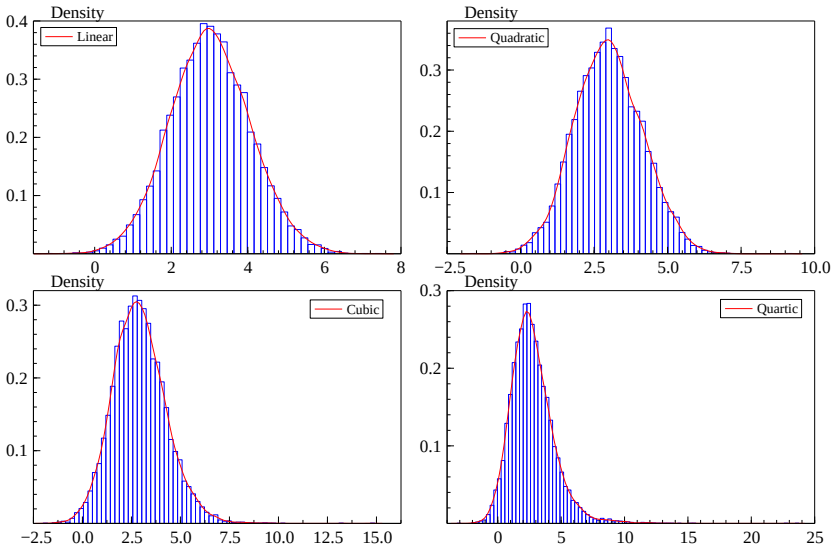

Figure 3.1: Distribution of t-statistics for a linear, quadratic, cubic, and quartic regressor under the alternative.

of the distribution of $\widehat{\sigma}_\beta^2$ under the alternative, which impacts on the distribution of the t-statistic. This is demonstrated in Figure 3.1, which records the distribution of the t-statistics for $\varphi_{\alpha,i} = 3$ and $T = 100$. Panels a–d correspond to the linear, quadratic, cubic and quartic DGPs respectively. The long right hand tail for the quartic is clearly evident. Edgeworth expansions would be a natural tool to approximate the small sample distribution, and we highlight this as a future area of research.

The resulting impact on the power of the optimal test, which is equivalent to a t-test of $\mathsf{H}_0 : \beta = 0$, is observed in Figure 3.2. The figure records the power of the optimal test based on the χ^2-approximation outlined in Section 3.2.4. Panels a and b record the powers for $T = 100$, and panels c and d record the corresponding powers for $T = 1,000$. For $T = 100$, the divergence between the analytic and Monte Carlo test powers is substantial, particularly for intermediate non-centralities, indicating that the χ^2-approximation is poor for higher order polynomials. Thus, caution must be applied when using conventional statistical analysis for processes that generate extreme observations, such as the cubic or quartic function.[2]

[2]The non-centralities were also calculated using the median of x_t^i as opposed to the mean. If the mean of x_t^i is large the results do deliver an improvement over the mean, but for a zero mean, this strategy fails badly. Furthermore, calculation of the mean based on the truncated distribution of x_t^i aimed to exclude the extreme observations causing the divergence. This is disastrous if the truncation is too severe, but does deliver marginal improvements if the truncation only excludes the few outliers far out in the tails: excluding observations that lie outside the 0.9995 quantile proved successful.

Table 3.1: Monte Carlo non-centralities for the optimal test

	$\varphi_{\alpha,i}$					
T	*1*	*2*	*3*	*4*	*5*	*6*
$i = 1$						
100	1.012	2.018	3.023	4.028	5.033	6.039
1,000	1.006	2.007	3.008	4.008	5.009	6.009
$i = 2$						
100	1.001	1.998	2.995	3.992	4.989	5.986
1,000	1.014	2.014	3.014	4.015	5.015	6.015
$i = 3$						
100	0.971	1.940	2.910	3.879	4.849	5.818
1,000	0.992	1.988	2.985	3.981	4.978	5.974
$i = 4$						
100	0.902	1.816	2.729	3.642	4.555	5.469
1,000	0.978	1.962	2.946	3.931	4.915	5.899

Notes: i = order of polynomial function: $i = 1$ is linear, $i = 2$ is quadratic, $i = 3$ is cubic and $i = 4$ is quartic.

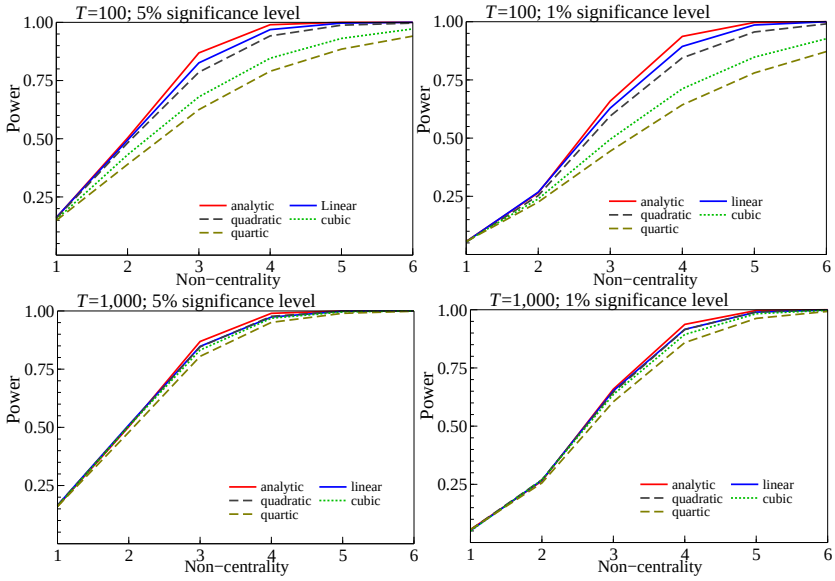

Figure 3.2: Analytic and Monte Carlo power functions for the optimal test on a polynomial function

3.2.7 Monte Carlo results

Monte Carlo experiments are undertaken to examine the power of the index test for static DGPs, for varying degrees of collinearity and number of regressors. Four tests are undertaken, including the optimal F-test on the known nonlinear functions (giving maximum power); White's test, which performs an F-test on the squares and cross-products of all the regressors; the index test, which computes the F-test for the orthogonalised quadratic functions; and the index test including orthogonalised cubic functions. We assess both polynomial DGPs and an LSTR specification.

Quadratic DGP

Consider the DGP given by:

$$y_t = \beta_1 x_{1,t} + \beta_2 x_{2,t} + \beta_3 x_{1,t}^2 + \epsilon_t, \tag{3.39}$$

where $\mathbf{x}_t = (x_{1,t}, x_{2,t})'$:

$$\mathbf{x}_t \sim \mathsf{IN}_2 [\mathbf{0}, \mathbf{\Omega}] \text{ with } \mathbf{\Omega} = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}, \tag{3.40}$$

and $\epsilon_t \sim \mathsf{IN}[0,1]$. As noted above, the optimal test of $\mathsf{H}_0 : \beta_3 = 0$, has a non-centrality of:

$$\varphi_{r,\alpha}^2 = 3T\beta^2, \tag{3.41}$$

which is independent of ρ, the degree of collinearity.

White's test is also independent of the degree of collinearity, and will have the same non-centrality as the optimal test but with larger degrees of freedom. The power of White's test for (3.39) will depend on:

$$\delta_1 x_{1,t}^2 + \delta_2 x_{2,t}^2 + \delta_3 x_{1,t} x_{2,t}. \tag{3.42}$$

The non-centrality of the White test is given by:

$$\begin{aligned} \varphi_{r,\alpha}^2 &= \frac{T\mathsf{E}\left[\left(\delta_1 x_{1,t}^2 + \delta_2 x_{2,t}^2 + \delta_3 x_{1,t} x_{2,t}\right)^2\right]}{\sigma_\epsilon^2} \\ &= T\left[3\delta_1^2 + 2\rho^2\delta_1\delta_2 + 12\rho\delta_1\delta_2 + 3\delta_2^2 + 12\rho\delta_2\delta_3 + \rho^2\delta_3^2\right]. \end{aligned} \tag{3.43}$$

using:[3]

$$\begin{aligned}\mathsf{E}\left[x_{i,t}^2 x_{j,t}^2\right] &= \mathsf{E}\left[\left(x_{i,t}x_{j,t}\right)^2\right] = \rho^2 \\ \mathsf{E}\left[x_{i,t}^3 x_{j,t}\right] &= 2\left[3\rho\right] = 6\rho.\end{aligned}$$

As $\delta_2 = \delta_3 = 0$, the test non-centrality collapses to the optimal test non-centrality of $3T\delta_1^2$. Hence, the difference between the analytic power of the optimal test and White's test will be a function of the number of degrees of freedom alone.

The index test for (3.39) is based on $\gamma_1 z_{1,t}^2 + \gamma_2 z_{2,t}^2$, but as:

$$\mathbf{z}_t^* = \mathbf{H}'\mathbf{x}_t$$

we can calculate:

$$z_{1,t} = x_{1,t} + \kappa_1 x_{2,t} \tag{3.44}$$

$$z_{2,t} = x_{2,t} + \kappa_2 x_{1,t} \tag{3.45}$$

where $\boldsymbol{\kappa}$ is based on the eigenvalues $\boldsymbol{\lambda}$, of $\boldsymbol{\Omega}$ (assuming a zero mean, the $z_{i,t}$ are given by $(\mathbf{H}'\mathbf{x}_t)\,\boldsymbol{\lambda}^{-\frac{1}{2}}$). Note that even under orthogonality, as the index is based on $\widehat{\boldsymbol{\Omega}}$, sampling errors will result in a non-zero off-diagonal. If $\boldsymbol{\Omega} = \mathbf{I}_2$, this implies $\boldsymbol{\lambda} = \mathbf{1}$, and $\mathbf{H} = \mathbf{I}_2$, such that $z_{1,t}^2 = x_{1,t}^2$ and $z_{2,t}^2 = x_{2,t}^2$. In practice, $\widehat{\boldsymbol{\Omega}} \neq \mathbf{I}_2$ such that $\mathbf{z}$ comprises a linear combination of the $\mathbf{x}$s.

This results in:

$$z_{1,t}^2 = x_{1,t}^2 + \kappa_1^2 x_{2,t}^2 + 2\kappa_1 x_{1,t}x_{2,t} \tag{3.46}$$

$$z_{2,t}^2 = x_{2,t}^2 + \kappa_2^2 x_{1,t}^2 + 2\kappa_2 x_{1,t}x_{2,t} \tag{3.47}$$

such that the index test power will depend on:

$$\gamma_1\left(x_{1,t}^2 + \kappa_1^2 x_{2,t}^2 + 2\kappa_1 x_{1,t}x_{2,t}\right) + \gamma_2\left(x_{2,t}^2 + \kappa_2^2 x_{1,t}^2 + 2\kappa_2 x_{1,t}x_{2,t}\right) \tag{3.48}$$

[3] Using the fact that the fourth cumulant of a normal is 0, the Hannan (1970, p.23) theorem states:

$$\mathsf{E}\left[x_{1,t}^3 x_{2,t}\right] = \mathsf{E}\left[w_{1,t}w_{2,t}w_{3,t}w_{4,t}\right] = 2\left\{\mathsf{E}\left[w_{1,t}w_{2,t}\right]\mathsf{E}\left[w_{3,t}w_{4,t}\right] + \mathsf{E}\left[w_{1,t}w_{3,t}\right]\mathsf{E}\left[w_{2,t}w_{4,t}\right] + \mathsf{E}\left[w_{1,t}w_{4,t}\right]\mathsf{E}\left[w_{2,t}w_{3,t}\right]\right\}.$$

Table 3.2: Analytic power for the nonlinearity tests

β	*0.1*	*0.15*	*0.2*	*0.25*	*0.3*	*0.35*
$\varphi^2_{r,\alpha}$	*3*	*6.75*	*12*	*18.75*	*27*	*36.75*
5%						
Optimal (1)	0.392	0.745	0.952	0.997	1.000	1.000
White (3)	0.135	0.403	0.750	0.951	0.996	1.000
Index (2)	0.221	0.547	0.858	0.981	0.999	1.000
1%						
Optimal (1)	0.186	0.493	0.822	0.973	0.999	1.000
White (3)	0.051	0.209	0.524	0.839	0.975	0.999
Index (2)	0.092	0.313	0.660	0.914	0.991	1.000

Notes: (.) denotes the degrees of freedom.

which can be solved using:

$$\gamma_1 + \gamma_2\kappa_2^2 \approx \beta \tag{3.49}$$

$$\gamma_1\kappa_1^2 + \gamma_2 \approx 0 \tag{3.50}$$

$$2\gamma_1\kappa_1 + 2\gamma_2\kappa_2 \approx 0. \tag{3.51}$$

Under orthogonality, the analytic non-centrality collapses to that of the optimal test, $3T\gamma_1^2$, but collinearity impacts on the non-centrality via the $\boldsymbol{\kappa}$ weighting.

Analytic powers of the tests for the bivariate case under orthogonality are reported in Table 3.2, where the difference in power arises due to differing degrees of freedom. The optimal test has 1 degree of freedom, the index test has 2 degrees of freedom and White's test has 3 degrees of freedom. While the index test should outperform White's test under orthogonality for (3.39) as $z_{i,t}^2 = x_{i,t}^2$ and there is one less degree of freedom, in practice the sampling error in $\widehat{\mathbf{\Omega}}$ results in a linear combination such that $z_{1,t}^2$ is contaminated by $x_{2,t}$, reducing the power to detect $x_{1,t}^2$. Thus, there is a trade-off between sampling error in the index test and degrees of freedom gains.

To investigate this impact we undertake a Monte Carlo experiment in which up to twelve white noise processes are generated, with $x_{i,t} \sim \mathsf{IN}[0,1]$, $\forall i = 1, ..., 12$, and $\mathsf{cov}[x_{i,t}x_{j,t}] = \rho$, $\forall i \neq j$, where $\rho = 0$ or 0.9.[4] n is the number of linear regressors, increasing from two to twelve in the GUM, of which only two ($x_{1,t}$ and $x_{2,t}$) are relevant.

The four tests undertaken are outlined below:

[4]Experiments for an intermediate range of ρ were also undertaken. Results are excluded from the figures for a clear exposition, but are available on request.

1. Optimal test, $\mathsf{H}_0 : \widetilde{\phi} = 0$ for:

$$y_t = \sum_{i=1}^{n} \phi_i x_{i,t} + \widetilde{\phi} x_{1,t}^2 + v_t. \tag{3.52}$$

2. White's test, $\mathsf{H}_0 : \widetilde{\boldsymbol{\delta}} = \mathbf{0}$ for:

$$y_t = \sum_{i=1}^{n} \delta_i x_{i,t} + \sum_{j=1}^{n} \sum_{k=1}^{n} \widetilde{\delta}_{jk} x_{j,t} x_{k,t} + u_t. \tag{3.53}$$

3. Index test, $\mathsf{H}_0 : \widetilde{\boldsymbol{\gamma}} = \mathbf{0}$ for:

$$y_t = \sum_{i=1}^{n} \gamma_i x_{i,t} + \sum_{j=1}^{n} \widetilde{\gamma}_j z_{j,t}^2 + e_t. \tag{3.54}$$

4. Index test including cubics, $\mathsf{H}_0 : \widetilde{\boldsymbol{\psi}} = \widetilde{\boldsymbol{\theta}} = \mathbf{0}$ for:

$$y_t = \sum_{i=1}^{n} \psi_i x_{i,t} + \sum_{j=1}^{n} \widetilde{\psi}_j z_{j,t}^2 + \sum_{k=1}^{n} \widetilde{\theta}_k z_{k,t}^3 + \eta_t. \tag{3.55}$$

Two sample sizes are considered, $T = 100$ and 300, and $M = 10,000$ replications are undertaken. We set $\beta_1 = \beta_2 = 0.3$ and $\beta_3 = 0.1732$ in the DGP (3.39), which results in non-centralities of three for all regressors under orthogonality, for $T = 100$. The coefficients are fixed for $T = 300$ to assess the impact of increasing sample size. The results are reported in Figure 3.3, where the number of linear regressors is recorded along the horizontal axis and power is recorded on the vertical axis. The top two panels record results for $T = 100$, and the bottom two panels record results for $T = 300$. The divergence between the analytic and optimal test is evident, as discussed above. Both the White and index test have a similar power for $n = 2$, but White's power declines rapidly as n increases because the test's degrees of freedom increase by $\frac{n(n+1)}{2}$. The decline in the index test power as n increases is not as sharp, although the degrees of freedom increase by n. The index test including cubic terms does have a lower power than the index test on the squares alone, although the magnitude of the loss in power is not substantial and is fairly constant across n and ρ. The index test has a marginally higher power under collinearity than orthogonality, which is counter-intuitive. This is due to the interactions between sampling error and collinearity, such that if the test was trying to detect $x_{1,t}^2$,

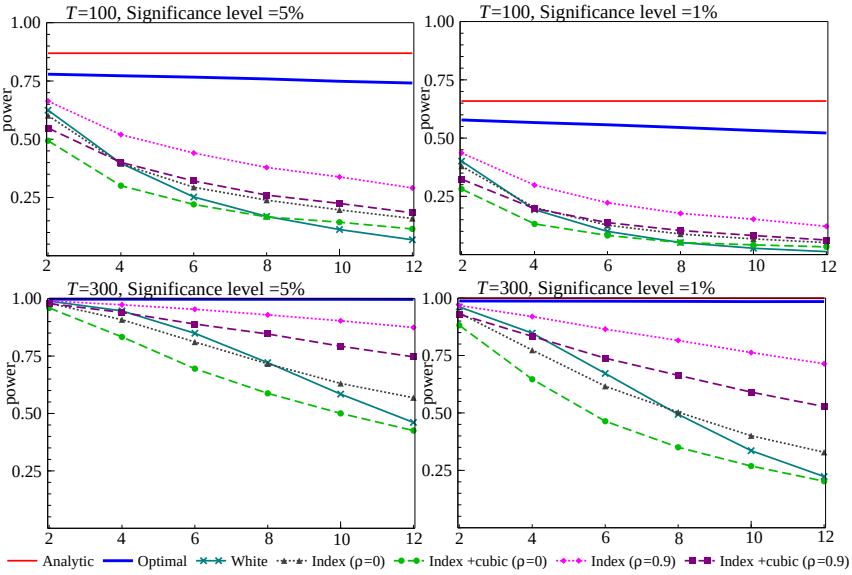

Figure 3.3: Power of nonlinearity tests for a quadratic function

but under orthogonality, $z_{1,t}^2$ contained some component of $x_{2,t}^2$ as $\widehat{\mathbf{\Omega}} \neq \mathbf{\Omega}$, a high ρ would increase the power to detect nonlinearity via $x_{2,t}^2$. White's test should, in theory, have a higher power than the collinear index test, but the costs in terms of the extra degrees of freedom imply that the index test does have marginally higher power for $\rho = 0.9$ when $n = 2$. As T increases to 300, the power of all tests increase, with a unit power for the analytic test.

Quadratic DGP including a cross-product

We next consider augmenting the DGP with a cross-product term, given by:

$$y_t = \beta_1 x_{1,t} + \beta_2 x_{2,t} + \beta_3 x_{1,t}^2 + \beta_4 x_{1,t} x_{2,t} + \epsilon_t, \tag{3.56}$$

where $\epsilon_t \sim \mathsf{IN}[0,1]$ and $\mathbf{x}_t$ is generated by (3.40). We set $\beta_1 = \beta_2 = \beta_4 = 0.2$ and $\beta_3 = 0.1155$, to result in a non-centrality of two for the individual regressors, for $T = 100$ under orthogonality. A lower non-centrality is specified because the DGP with two non-linear terms results in a high probability of detection. The tests are as outlined above, with the optimal test given by $\mathsf{H}_0 : \widetilde{\phi}_1 = \widetilde{\phi}_2 = 0$ for:

$$y_t = \sum_{i=1}^{n} \phi_i x_{i,t} + \widetilde{\phi}_1 x_{1,t}^2 + \widetilde{\phi}_2 x_{1,t} x_{2,t} + v_t. \tag{3.57}$$

Figure 3.4 records the results for $T = 100$, with the top two panels corresponding

to $\rho = 0$, and the bottom two panels corresponding to $\rho = 0.9$ for the same parameter values. As the power of the optimal test depends on:

$$\beta_3 x_{1,t}^2 + \beta_4 x_{1,t} x_{2,t},$$

under collinearity, the non-centrality of the optimal F-test is:

$$\begin{aligned} \varphi_{\alpha,i}^2 &= \frac{T\mathsf{E}\left[\left(\beta_3 x_{1,t}^2 + \beta_4 x_{1,t} x_{2,t}\right)^2\right]}{\sigma_\epsilon^2} \\ &= T\left[\beta_3^2 x_{1,t}^4 + \beta_4^2 x_{1,t}^2 x_{2,t}^2 + 2\beta_3\beta_4 x_{1,t}^3 x_{2,t}\right] \\ &= T\left[3\beta_3^2 + \beta_4^2\rho^2 + 12\beta_3\beta_4\rho\right]. \end{aligned} \tag{3.58}$$

Hence, the non-centrality of the joint F-test is $\varphi_{r,\alpha} = 5.7$ for $\rho = 0.9$, delivering a much higher power for all tests under collinearity. Note that the correlation between the linear and nonlinear regressors is zero regardless of ρ:

$$\mathsf{E}\left[\left(\beta_1 x_{1,t} + \beta_2 x_{2,t}\right)\left(\beta_3 x_{1,t}^2 + \beta_4 x_{1,t} x_{2,t}\right)\right] = 0,$$

as:

$$\begin{aligned} \mathsf{E}\left[x_{1,t} x_{1,t}^2\right] &= \mathsf{E}\left[x_{1,t}^3\right] = 0 \\ \mathsf{E}\left[x_{1,t}^2 x_{2,t}\right] &= \mathsf{E}\left[x_{1,t} x_{2,t}^2\right] = 0 \end{aligned}$$

due to zero skewness under Gaussianity.[5]

Under orthogonality, for $n = 2$, including $\sum_{j=1}^2 \alpha_j z_{j,t}^2$ will just retain $x_{1,t}^2 + x_{2,t}^2$, as $\mathbf{\Omega}$ implies $\mathbf{H} = \mathbf{I}_2$, and will exclude combinations of $x_{1,t}$ and $x_{2,t}$. Hence, the test must have low power against the cross-product term in (3.56). However, as the index is based on $\widehat{\mathbf{\Omega}}$, sampling errors will result in a non-zero off-diagonal to result in (3.46) and (3.47). For the test to have power against (3.56), we require a low weight on the $x_{2,t}^2$ term in the $z_{1,t}^2$ equation. Consider a weighted average of the regressors:

$$z_{1,t}^2 + \delta z_{2,t}^2 = \left(1 + \delta\kappa_2^2\right) x_{1,t}^2 + \left(\delta + \kappa_1^2\right) x_{2,t}^2 + 2\left(\kappa_1 + \delta\kappa_2\right) x_{1,t} x_{2,t}. \tag{3.59}$$

Hence, a negative δ would downweight the $x_{2,t}^2$, yielding the highest power for the index test: regression should select such a weighting. Nevertheless, there is no close approxi-

[5] As $x_{1,t}^2 > 0$, whereas $x_{2,t}$ is symmetric about zero, their correlation will be zero.

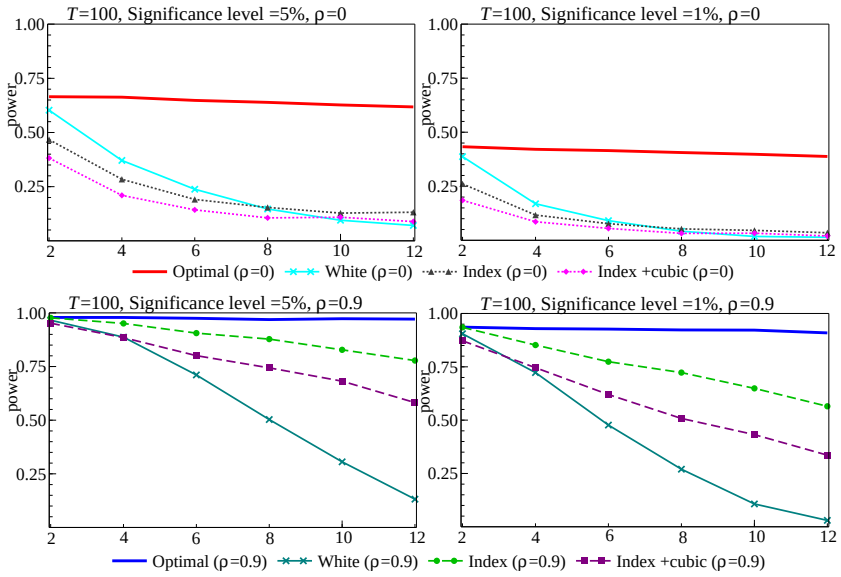

Figure 3.4: Power of nonlinearity tests for a quadratic and cross-product function

mation to $\beta_3 x_{1,t}^2 + \beta_4 x_{1,t} x_{2,t}$ since $\delta = -\kappa_1^2$ yields (normalising on $z_{1,t}^2$):

$$z_{1,t}^2 - \kappa_1^2 z_{2,t}^2 = \left(1 - \kappa_1^2 \kappa_2^2\right) x_{1,t}^2 + 2\kappa_1 \left(1 - \kappa_1 \kappa_2\right) x_{1,t} x_{2,t} \simeq x_{1,t}^2 + 2\kappa_1 x_{1,t} x_{2,t}, \tag{3.60}$$

which will be close only when $\beta_4 = 2\kappa_1 \beta_3$ for small κ_1.

Under orthogonality, White's test has a higher power than the index test for small n, with the degrees of freedom advantages of the index test only resulting in a higher power when n is larger than 8, because the index test has low power to detect the cross-product. Under collinearity, the index test has a much higher power, outperforming White's test for all n.

Cubic DGP

We next consider the case in which a cubic term is included in the DGP:

$$y_t = \beta_1 x_{1,t} + \beta_2 x_{2,t} + \beta_3 x_{1,t}^3 + \epsilon_t, \quad \epsilon_t \sim \mathsf{IN}\left[0, 1\right]. \tag{3.61}$$

To calculate the non-centralities under orthogonality, such that we can analyse $x_{2,t}$ independently, we shall use partitioning (see, e.g., Hendry, 1995, p.700). If we define $\mathbf{Z} = (x_{1,t}, x_{1,t}^3)'$ where:

$$\mathbf{Q} = \mathbf{I} - \mathbf{x}_1 \left(\mathbf{x}_1' \mathbf{x}_1\right)^{-1} \mathbf{x}_1',$$

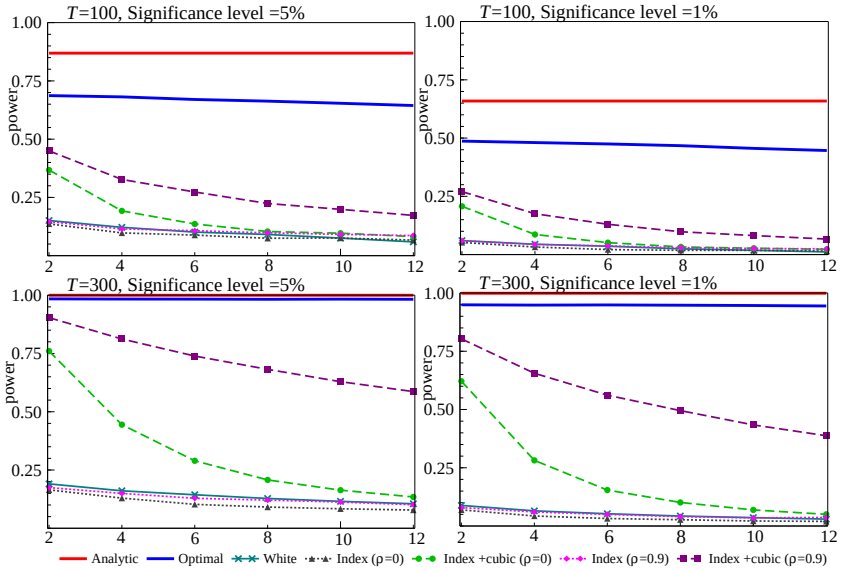

Figure 3.5: Power of nonlinearity tests for a cubic function

this implies:

$$\varphi^2_{r,\alpha}(\beta_3) = \frac{\beta_3^2 \mathsf{E}\left[\mathbf{x}_1^{3\prime}\mathbf{Q}\mathbf{x}_1^3\right]}{\sigma_\epsilon^2} = 6T\beta_3^2. \tag{3.62}$$

For the converse:

$$\varphi^2_{r,\alpha}(\beta_1) = \frac{\beta_1^2 \mathsf{E}\left[\mathbf{x}_1'\widetilde{\mathbf{Q}}\mathbf{x}_1\right]}{\sigma_\epsilon^2} = 0.4T\beta_1^2, \tag{3.63}$$

where $\widetilde{\mathbf{Q}} = \mathbf{I} - \mathbf{x}_1^3\left(\mathbf{x}_1^{3\prime}\mathbf{x}_1^3\right)^{-1}\mathbf{x}_1^{3\prime}$. Under orthogonality, the non-centrality of β_2 is as standard:

$$\varphi^2_{r,\alpha}(\beta_2) = T\beta_2^2. \tag{3.64}$$

For the Monte Carlo experiments we shall set $\beta_1 = 0.4743$, $\beta_2 = 0.3$ and $\beta_3 = 0.1225$, which results in a non-centrality of three for all regressors for $T = 100$ under orthogonality. We hold the coefficients constant to assess the impact of increasing the sample size to $T = 300$.

Figure 3.5 records the powers for $T = 100$ and 300 over a range of n. The results demonstrate that the cost of including the cubic term in the index, in terms of degrees of freedom, is small compared to the loss in power if the cubic terms are excluded, as this ensures against nonlinearity that manifests itself as a cubic type of function. Both the index test excluding cubic terms and White's test have similar powers, and there is no benefit to using the index test as n increases. The index test including cubic terms

does decline in power as n increases, suggesting that there are benefits to specifying a parsimonious model. A high degree of collinearity is again beneficial. This is because $\rho \neq 0$ results in combinations of $x_{1,t}$ and $x_{2,t}$ that proxy $x_{1,t}^3$. The linear combination of the $\mathbf{x}_t$s for $\mathbf{z}_t$ is:

$$z_{1,t}^3 = x_{1,t}^3 + 3\kappa_1 x_{1,t}^2 x_{2,t} + 3\kappa_1^2 x_{1,t} x_{2,t}^2 + \kappa_1^3 x_{2,t}^3 \tag{3.65}$$

$$z_{2,t}^3 = x_{2,t}^3 + 3\kappa_2 x_{2,t}^2 x_{1,t} + 3\kappa_2^2 x_{2,t} x_{1,t}^2 + \kappa_2^3 x_{1,t}^3 \tag{3.66}$$

Therefore, if $\rho = 0.9$, the index test will gain power to detect $x_{1,t}^3$ via the linear combinations, $x_{1,t}^2 x_{2,t}$ and $x_{1,t} x_{2,t}^2$. The gap between the analytic and optimal test is larger than for the quadratic DGP, in keeping with our previous analysis. Furthermore, for small n, the index test including cubic terms has a power close to that of the optimal, especially at $T = 300$, despite the test including all the irrelevant quadratic terms as well.

Quartic DGP

While a quartic function is somewhat extreme, and the small sample distribution of the t-statistic is poor, we argue that the index test based on quadratic functions should have power against quartic functions due to the collinearity between the two. Hence, we assess a quartic DGP given by:

$$y_t = \beta_1 x_{1,t} + \beta_2 x_{2,t} + \beta_3 x_{1,t}^4 + \eta_t, \tag{3.67}$$

where $\beta_1 = \beta_2 = 0.3$ and $\beta_3 = 0.0293$, which delivers a non-centrality of three for all regressors under orthogonality, for $T = 100$.

The results are recorded in Figure 3.6, where the substantial gap between the analytic and optimal test is evident. While the power is lower than that for the quadratic function, both White's test and the index test do have power against a quartic function. The patterns exhibited by the power functions correspond to those for the quadratic function, although the power does not decline to the same extent as n increases. Including the cubic term in the index test does not improve the power as there is no correlation between the third and fourth order polynomials.

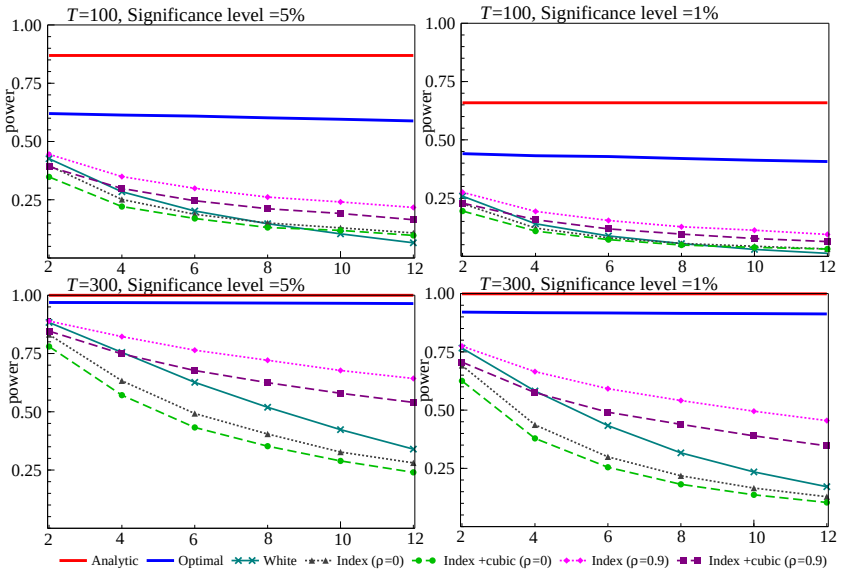

Figure 3.6: Power of nonlinearity tests for a quartic function

LSTR model

There are many potential feasible nonlinear functional forms. As the index test is designed to be used against a general alternative, it is important to see whether the test has power against specific functional forms. We shall focus on the LSTR model, as this model is examined in section 3.3.2. The Monte Carlo is necessarily equation specific, but it is indicative of the performance of the index test against White's test. The DGP is given as:

$$y_t = \beta_0 + \beta_1 x_{1,t} + \beta_2 x_{2,t} + (\delta_0 + \delta_1 x_{1,t} + \delta_2 x_{2,t}) \left[1 + \exp\left(-\gamma \left(x_{1,t} - c\right)\right)\right]^{-1} + \epsilon_t, \quad (3.68)$$

where $\epsilon_t \sim \mathsf{IN}[0,1]$ and the $\mathbf{x}_t$ process is generated by (3.40). Two parameterisations are considered, given by $\beta_0 = 0.2$, $\delta_0 = 0.8$, $\beta_1 = \delta_1 = 0.3$, $\beta_2 = \delta_2 = 0.4$, $\gamma = 2.5$, and $c = 0.5$; and alternatively, $\beta_0 = 0$, $\delta_0 = 0.2$, $\beta_1 = \beta_2 = 0.2$, $\delta_1 = \delta_2 = 0.4$, $\gamma = 2$, and $c = 0.1$. The first parameterisation delivers 'strong nonlinearity', in the sense that the two regimes are clearly distinct. The second parameterisation delivers 'weak nonlinearity', as the two regimes are closer in mean and the transition function is less rapid. $T = 100$, and is increased to $T = 200$ and $T = 300$ for the 'strong' and 'weak' cases respectively, for $M = 10,000$ replications.

While the optimal infeasible test based on the LSTR specification is not computed,

we do compute the power of the feasible test based on the known Taylor approximation (discussed in detail in section 3.3.2). We can replace the transition function by a third order Taylor expansion:

$$[1+\exp\{-\gamma(x_{1,t}-c)\}]^{-1} \simeq \frac{1}{2}+\frac{\gamma(x_{1,t}-c)}{4}-\frac{(\gamma(x_{1,t}-c))^3}{48}, \tag{3.69}$$

such that the approximation of (3.68) has the specification:

$$y_t \simeq \theta_0+\theta_1 x_{1,t}+\theta_2 x_{2,t}+\theta_3 x_{1,t}^2+\theta_4 x_{1,t}^3+\theta_5 x_{1,t}^4+\theta_6 x_{1,t}x_{2,t}+\theta_7 x_{1,t}^2 x_{2,t}+\theta_8 x_{1,t}^3 x_{2,t}+\varepsilon_t. \tag{3.70}$$

Hence, the Taylor approximation test is highly parameterised, and there may be degrees of freedom gains from the index test.

The results are recorded in Figures 3.7 and 3.8 for the two alternative DGPs. The figures record the Taylor approximation test based on (3.70), White's test and the index test for $\rho = 0$ and $\rho = 0.9$. The non-centralities of all tests differ under $\rho \neq 0$. All tests have high power under 'strong nonlinearity', indicating that polynomial approximations do capture nonlinearities that are generated by a smooth transition model. The power under collinearity is substantially, and consistently, higher than the power under orthogonality for all tests. The index test including cubic terms does not deliver a higher power than just including the quadratic terms, which is surprising given the Taylor approximation specification. Under weaker nonlinearity, the power function for White's test exhibits much higher power for $T = 300$ when $\rho = 0.9$ than would be anticipated given the results for $T = 100$. Alternative LSTR specifications are needed to draw more rigorous conclusions.

To conclude, we find that White's test and the index test perform comparatively for small n, but the power of White's test declines more sharply as n increases due to the rapidly increasing degrees of freedom. A preferable test may be to use White's test for $T >> n^2$ and small n, and then switch to the index test as $\frac{n^2}{T} > k$, where k is some threshold value such as 0.25. Furthermore, while parsimony delivers a higher power, such that selection prior to implementing the test would appear to be beneficial, this may be a hazardous strategy if the linear term is irrelevant but it enters the DGP in a nonlinear

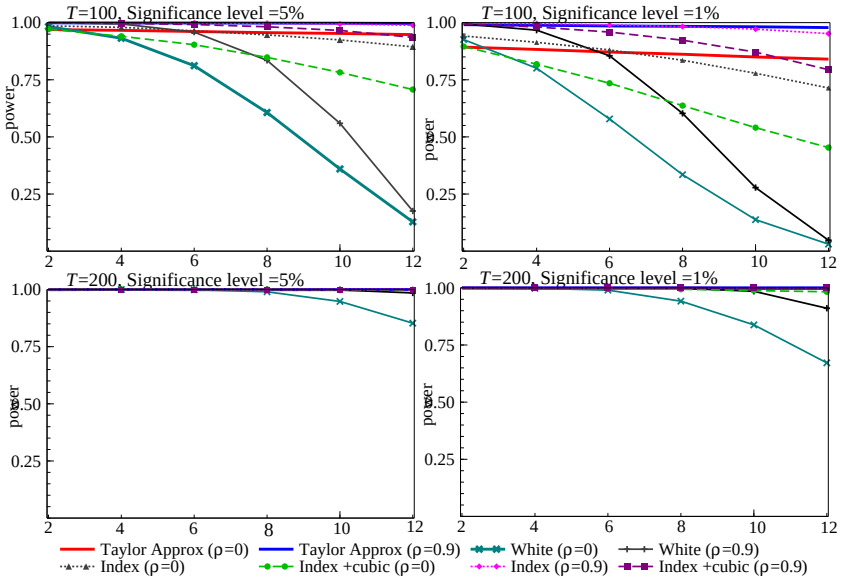

Figure 3.7: Power of nonlinearity tests for an LSTR function: strong nonlinearity

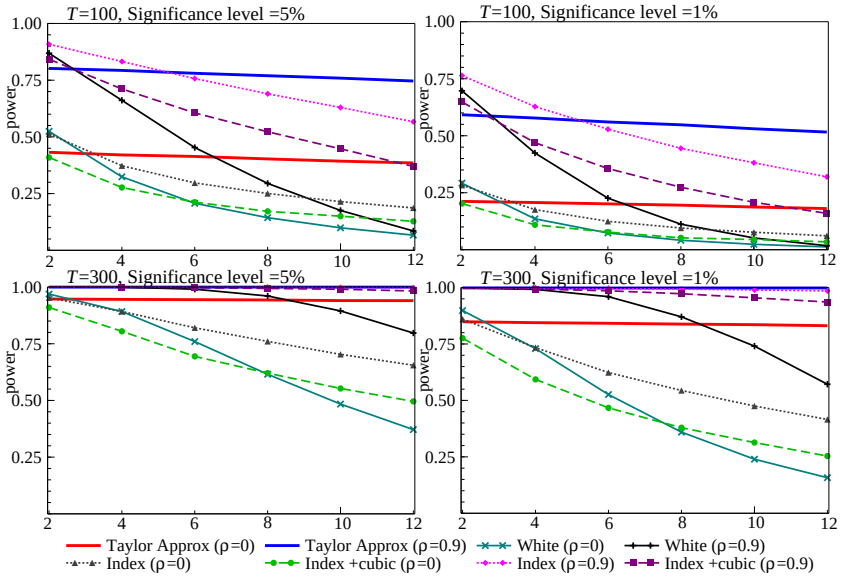

Figure 3.8: Power of nonlinearity tests for an LSTR function: weak nonlinearity

function. Thus, there is a trade-off between a higher power after selection and a risk of eliminating variables that are relevant via a nonlinear transformation, resulting in a lower power to detect nonlinearity if such a variable is excluded.

When the functional form and the set of relevant variables are both unknown, but nest the LDGP, the index test using quadratic and cubic terms has power to reject a false null in a wide range of circumstances: pure quadratic, pure cubic, pure quartic, and these in combinations, even for highly collinear data. It outperforms White's test in most of these situations, and can even be close to the optimal infeasible test. For larger departures from linearity, where several nonlinear terms occur for several variables, its power will dominate that in the experiments illustrated here. Thus, it promises to be a useful mis-specification test prior to undertaking nonlinear model selection.

3.2.8 Inverse uniform random variables

One class of functions that the test may not have power against are inverses. These often play a role in economic theory and have been implemented, for example, in RETINA, as a possible class of nonlinear functions. We consider the inverse of uniform random numbers (standard normal inverses are problematic as extreme realisations can be drawn with a mean zero process) to determine whether polynomial functions will capture this nonlinearity. To do this, we derive the correlations between a uniform variable and its inverse, as well as the correlations between a quadratic uniform and its inverse. If the LDGP contains an inverse polynomial function, we demonstrate that the polynomial will detect this form of nonlinearity as there is a high correlation between a variable and its inverse.

Consider:

$$x_t \sim \mathsf{IU}\left[a-h, a+h\right], \tag{3.71}$$

so:

$$\int_{(a-h)}^{(a+h)} f\left(x\right) dx = \frac{1}{2h}\int_{(a-h)}^{(a+h)} dx = \frac{1}{2h}\left[x\right]_{(a-h)}^{(a+h)} = \frac{1}{2h}\left((a+h)-(a-h)\right) = 1, \tag{3.72}$$

where $\mathsf{E}[x_t] = a$ and $\mathsf{V}[x_t] = h^2/3$. The density of a uniform is given by $f\left(x\right) = \left(2h\right)^{-1}$.

Let $y = g(x) = x^{-1}$ so $x = g^{-1}(y) = y^{-1}$ and:

$$\left|\frac{\partial x}{\partial y}\right| = \left|\frac{\partial y^{-1}}{\partial y}\right| = \frac{1}{y^2},$$

confirmed by:

$$\frac{1}{2h}\int_{(a+h)^{-1}}^{(a-h)^{-1}} \frac{1}{y^2}dy = \frac{1}{2h}\left[-y^{-1}\right]_{(a+h)^{-1}}^{(a-h)^{-1}} = \frac{1}{2h}\left(-(a-h)+(a+h)\right) = \frac{1}{2h}(2h) = 1.$$

We next calculate the moments:

$$\begin{aligned}
\mathsf{E}[y_t] &= \frac{1}{2h}\int_{(a+h)^{-1}}^{(a-h)^{-1}} y^{-1}dy = \frac{1}{2h}\left[\log(y)\right]_{(a+h)^{-1}}^{(a-h)^{-1}} = \frac{1}{2h}\left(\log(a+h)-\log(a-h)\right) \\
&\simeq \frac{1}{2h}\left(\ln a + \frac{1}{a}h - \frac{1}{2a^2}h^2 + \frac{1}{3a^3}h^3 - \ln a + \frac{1}{a}h + \frac{1}{2a^2}h^2 + \frac{1}{3a^3}h^3\right) \\
&= \frac{1}{a} + \frac{h^2}{3a^3} = \left(\frac{3a^2+h^2}{3a^3}\right) \qquad (3.73)
\end{aligned}$$

and:

$$\begin{aligned}
\mathsf{V}[y_t] &= \frac{1}{2h}\int_{(a+h)^{-1}}^{(a-h)^{-1}} \left[1 - \frac{2}{y}\left(\frac{1}{2h}\left(\log(a+h)-\log(a-h)\right)\right)\right. \\
&\qquad \left. + \frac{\left(\frac{1}{2h}\left(\log(a+h)-\log(a-h)\right)\right)^2}{y^2}\right]dy \\
&= \frac{1}{2h}\int_{(a+h)^{-1}}^{(a-h)^{-1}} dy - \frac{2}{2h}\left(\frac{1}{2h}\left(\log(a+h)-\log(a-h)\right)\right)\int_{(a+h)^{-1}}^{(a-h)^{-1}} \frac{1}{y}dy \\
&\quad + \frac{1}{2h}\left(\frac{1}{2h}\left(\log(a+h)-\log(a-h)\right)\right)^2 \int_{(a+h)^{-1}}^{(a-h)^{-1}} \frac{1}{y^2}dy \\
&= \frac{1}{2h}\left((a-h)^{-1} - (a+h)^{-1}\right) - \left(\frac{1}{2h}\left(\log(a+h)-\log(a-h)\right)\right)^2. \qquad (3.74)
\end{aligned}$$

Finally we can calculate the correlation coefficient between the uniform variable x_t and its inverse y_t as:

$$\begin{aligned}
\rho_{x,y}^2 &= \frac{\left(\mathsf{E}\left[\left(x_t - \mathsf{E}[x_t]\right)\left(y_t - \mathsf{E}[y_t]\right)\right]\right)^2}{\mathsf{V}[x_t]\,\mathsf{V}[y_t]} = \frac{\left(\mathsf{E}[x_t y_t - \mathsf{E}[x_t]\,\mathsf{E}[y_t]]\right)^2}{\mathsf{V}[x_t]\,\mathsf{V}[y_t]} = \frac{\left(1 - \mathsf{E}[x_t]\,\mathsf{E}[y_t]\right)^2}{\mathsf{V}[x_t]\,\mathsf{V}[y_t]} \\
&= \frac{3\left(1 - a\left(\frac{1}{2h}\left(\log(a+h)-\log(a-h)\right)\right)\right)^2}{h^2\left(\frac{1}{2h}\left((a-h)^{-1} - (a+h)^{-1}\right) - \left(\frac{1}{2h}\left(\log(a+h)-\log(a-h)\right)\right)^2\right)}. \qquad (3.75)
\end{aligned}$$

Undertaking the same analysis for the quadratic transformation given by:

$$w_t = k(x) = x_t^2$$

where x_t is given by (3.71), and $x = k^{-1}(w) = \sqrt{w}$, such that:

$$\left|\frac{\partial x}{\partial w}\right| = \left|\frac{\partial w^{\frac{1}{2}}}{\partial w}\right| = \frac{1}{2}w^{-\frac{1}{2}},$$

the moments are given by:

$$\begin{aligned}\mathsf{E}[w_t] &= \frac{1}{2h}\int_{(a-h)^2}^{(a+h)^2}\frac{1}{2}w^{\frac{1}{2}}dw = \frac{1}{6h}\left[w^{\frac{3}{2}}\right]_{a-h^2}^{a+h^2} \\ &= \frac{1}{6h}\left[6a^2h + 2h^3\right] = a^2 + \frac{h^2}{3} \end{aligned} \tag{3.76}$$

and:

$$\begin{aligned}\mathsf{V}[w_t] &= \frac{1}{2h}\int_{(a-h)^2}^{(a+h)^2}\left[w - \left(a^2 + \frac{h^2}{3}\right)\right]^2 w^{-\frac{1}{2}}dw \\ &= \frac{1}{4h}\int_{(a-h)^2}^{(a+h)^2} w^2dw + \frac{a^2 + \frac{h^2}{3}}{2h}\int_{(a+h)^2}^{(a-h)^2} w^{\frac{1}{2}}dw + \left(a^2 + \frac{h^2}{3}\right)^2\int_{(a-h)^2}^{(a+h)^2} w^{-\frac{1}{2}}dw \\ &= \frac{1}{4h}\left[\frac{2}{5}\left((a+h)^5 - (a-h)^5\right) - \frac{4}{3}\left((a+h)^3 - (a-h)^3\right)\left(a^2 + \frac{h^2}{3}\right)\right. \\ &\qquad \left. +4h\left(a^2 + \frac{h^2}{3}\right)^2\right]. \end{aligned} \tag{3.77}$$

We also compute the moments for the inverse quadratic function in order to obtain the correlation coefficient. Let $z = h(x) = x^{-2}$, where x_t is given by (3.71), so $h^{-1}(z) = z^{-\frac{1}{2}}$ and $\left|\frac{\partial x}{\partial z}\right| = \left|\frac{\partial z^{-\frac{1}{2}}}{\partial z}\right| = -\frac{1}{2}y^{-\frac{3}{2}}$. Then:

$$\begin{aligned}\mathsf{E}[z_t] &= \frac{1}{4h}\int_{(a+h)^{-2}}^{(a-h)^{-2}} z^{\frac{1}{2}}dz = \frac{1}{2h}\left[z^{\frac{1}{2}}\right]_{(a+h)^{-2}}^{(a-h)^{-2}} \\ &= \frac{1}{2h}\left[(a-h)^{-1} - (a+h)^{-1}\right] \end{aligned} \tag{3.78}$$

and:

$$\begin{aligned}\mathsf{V}[z_t] &= \frac{1}{4h}\int_{(a+h)^{-2}}^{(a-h)^{-2}}\left(z - \frac{1}{2h}\left[(a-h)^{-1} - (a+h)^{-1}\right]\right)^2 z^{-\frac{3}{2}}dz \\ &= \frac{1}{4h}\int_{(a+h)^{-2}}^{(a-h)^{-2}} z^{\frac{1}{2}}dz + \frac{(a-h)^{-1} - (a+h)^{-1}}{4h^2}\int_{(a-h)^{-2}}^{(a+h)^{-2}} z^{-\frac{1}{2}}dz \\ &\qquad + \frac{\left((a-h)^{-1} - (a+h)^{-1}\right)^2}{16h^3}\int_{(a+h)^{-2}}^{(a-h)^{-2}} z^{-\frac{3}{2}}dz \\ &= \frac{1}{4h}\left[\frac{2}{3}\left((a-h)^{-3} - (a+h)^{-3}\right) - \frac{1}{h}\left[(a-h)^{-1} - (a+h)^{-1}\right]^2\right]. \end{aligned} \tag{3.79}$$

Finally, we compute the correlation coefficient between x_t^2 and x_t^{-2} as:

$$\begin{aligned}\rho_{w,z}^2 &= \frac{(1-\mathsf{E}[w_t]\,\mathsf{E}[z_t])^2}{\mathsf{V}[w_t]\,\mathsf{V}[z_t]} \\ &= \frac{\left[1-\left(a^2+\frac{h^2}{3}\right)\left[\frac{1}{2h}\left((a-h)^{-1}-(a+h)^{-1}\right)\right]\right]^2}{\mathsf{V}[w_t]\,\mathsf{V}[z_t]}. \end{aligned} \tag{3.80}$$

Evaluating expressions (3.75) and (3.80) for $x_t \sim \mathsf{IU}[1,2]$, so $a = 1.5$ and $h = 0.5$, yields $|\rho_{x,y}| = 0.9842$ and $|\rho_{w,z}| = 0.9393$. Sample estimates based on $10,000$ draws yield $|\hat{\rho}_{x,y}| = 0.9841$ and $|\hat{\rho}_{w,z}| = 0.9392$, confirming these results. Thus, x_t is a close approximation to x_t^{-1} for this case; the interest of this result is that the linear term may 'pick up' inverses, which is not the index test. Furthermore, the quadratic term would pick up inverse quadratics, as x_t^2 is a close approximation to x_t^{-2}, although these will obviously not be uniform. Thus, the index test will have power against inverse polynomial functions of the type generated by RETINA, and this is yet another generalisation of the directions in which the index test will have power.

To conclude this section, we find that the index test has a high power (conditional on the non-centrality parameter) when the degree of collinearity is large and when the number of regressors is large, as the dimension reduction benefits are most substantial. The inclusion of the cubic index is recommended as the loss in power is small when the DGP is quadratic, but the increase in power for cubic DGPs is substantial. Given that the index test is easily implemented and exhibits excellent power properties for large dimensional systems, it provides a strong basis for examining the functional form of the specified GUM, and is recommended to be implemented in the nonlinear selection strategy.

3.3 Nonlinear functions

Econometric modelling of nonlinear processes presents many problems over and above those encountered when developing linear models. Identifying a unique nonlinear representation of an economic process can be formidable given the complexity of possible LDGPs. As there are an infinite number of potential functional forms that the LDGP may take, specifying a GUM that nests the unknown LDGP is problematic. The two

concerns for the econometrician include identification of the relevant variables and specification of the functional form.

The methodology suggested by the *Gets* framework is to explore, conditional on theory, sufficiently general models to nest a class of LDGPs. If the model does not match with the conjectured LDGP, the nonlinear GUM is revised to consider another class of nonlinear models, and the procedure continues iteratively. We focus on polynomials as they provide a good local approximation for a wide range of nonlinear models. Other advantages of the polynomial class are that simple operational rules for orthogonalising can be applied, and the class will retain linearity in the parameters. PcGets uses standard OLS or IV estimation, and we aim to incorporate a nonlinear capability within this framework, hence, our preference for linearity in the parameters. In principle, a general nonlinear, likelihood based, system approach is feasible, but no software yet exists to implement such a procedure. However, models that are nonlinear in the parameters are a result of restrictions that are imposed on the parameters, and the gains achieved from these more complex models are often limited compared to the costs of estimating such models within a *Gets* framework.

Other potential approximations that were considered include orthogonal polynomials such as Hermite polynomials, but these perform poorly in the tails; Fourier series, but these require a large number of terms to obtain a close approximation; and asymptotic series, but these tend to be intractable. Confluent hypergeometric functions provide a very general functional form that can capture a wide range of nonlinearities, see Abadir (1999), and this warrants further investigation.

3.3.1 Polynomial series

If the functional form of the nonlinear DGP is unknown, a Taylor series expansion delivers a polynomial specification, see (1.7), which motivates the use of polynomial functions.[6] While (1.7) is intractable, the inclusion of more variables than observations in regression

[6]Polynomial functions are commonly used in economics and are useful because of Weierstrass's approximation theorem, which states that any continuous function on a closed and bounded interval can be approximated by polynomials, i.e. if $x \in [a,b]$, for any $\epsilon > 0$ there exists a polynomial $p(x) \in [a,b]$ such that $|f(x) - P(x)| < \epsilon \; \forall x \in [a,b]$.

analysis is feasible, enabling a greater flexibility when examining nonlinear models as the number of potential regressors is likely to be large.

3.3.2 Application: Approximation to a LSTR model

STR models are a popular functional form for nonlinear econometric models, capturing a wide range of regime-switching theories. We undertake a Monte Carlo experiment to establish the ability of a Taylor expansion to capture the characteristics of a STR model, therefore providing support for the strategy of approximating unknown nonlinearity by a polynomial approximation.

A general STR model is outlined in (1.6), repeated here for clarity:

$$y_t = \boldsymbol{\beta}'\mathbf{X}_t + \left(\boldsymbol{\theta}'\mathbf{X}_t\right)\mathsf{G}\left(\gamma, c, s_t\right) + u_t, \quad u_t \sim \mathsf{IN}\left[0, \sigma_u^2\right], \tag{3.81}$$

for $t = 1, ..., T$, where $\mathsf{G}\left(.\right)$ is the transition function. Various distributional assumptions can be made on the transition function, and we investigate the properties of the LSTR(1) given by:

$$\mathsf{G}\left(\gamma, c, s_t\right) = \left[1 + \exp\left\{-\gamma\left(\frac{s_t - c}{\widehat{\sigma}_s}\right)\right\}\right]^{-1}. \tag{3.82}$$

In this monotonic transition function, γ determines how rapid the transition is from 0 to 1 as a function of the transition variable s_t, and c determines where the transition occurs. As $\gamma \to \infty$ the model becomes a two regime-switching regression model, and $\gamma > 0$ is the identifying restriction. Estimation of γ is particularly difficult as the likelihood function is not well behaved. It is common for the likelihood function of $\widehat{\gamma}$ to be fairly flat for a wide range of γ, particularly if γ is high, resulting in optimisation problems. Furthermore, a unique optimum may not exist for high γ. Extensive Monte Carlo simulations and empirical applications suggest an upper bound on $\widehat{\gamma}$ of approximately five.[7] For $\widehat{\gamma} \geq 5$, the transition function acts as a two regime-switching process, and a simplification to

[7] Chebyshev's inequality states that for any density function with finite first two moments:

$$\mathsf{P}\left(|X - \mathsf{E}\left(X\right)| \geq k\sigma_x\right) \leq \frac{1}{k^2}$$

such that deviations of more than $2\sigma_x$ have a maximum probability of 75%, and more extreme outcomes of $5\sigma_x$ will only be exceeded 4% of the time (see Hendry, 1995, p.666). Thus, defining $\mathsf{G}\left(\gamma, c, s_t\right) = \left[1 + e^{-\gamma s_t^*}\right]^{-1}$, where $s_t^* = \frac{s_t - c}{\widehat{\sigma}_s}$, we can assume $s_t^* \sim \mathsf{D}\left[0, 1\right]$, and therefore in theory, we could use Chebyshev's inequality to derive $\mathsf{P}\left(|\gamma s_t^*| \geq k\sigma\right) \leq \frac{1}{k^2\gamma^2}$ and solve for γ.

a switching regression model can be made. For values of $\widehat{\gamma}$ close to 0, the increased uncertainty regarding the regime increases the uncertainty of $\widehat{\boldsymbol{\theta}}$, and consequently the estimates of $\widehat{\boldsymbol{\beta}}$ that correspond to the variables that also interact with the transition function.

To provide empirical support for our supposition that $\widehat{\gamma} > 5$ can be equivalently modelled as a regime-switching model, we assess evidence from Bårdsen, Hurn and McHugh (2003a), who develop an LSTR(1) model of the Australian unemployment rate for the period 1980q4-2001q1 using a *Gets* modelling approach. The quarterly change in unemployment (recorded in Figure 3.9a) is modelled as a function of annual output growth, real unemployment benefits, aggregate labour productivity, and the average real wage. The transition function is given by:

$$\mathsf{G}(\gamma, c, s_t) = \left[1 + \exp\left\{-\gamma\left(\frac{\Delta_4 u_{t-1} - 0.165}{\widehat{\sigma}_{\Delta_4 u_{t-1}}}\right)\right\}\right]^{-1}, \tag{3.83}$$

where the transition variable is the lagged annual change in unemployment (recorded in Figure 3.9b). Figure 3.9, panels c and d, record the transition function for various γ. A high γ results in a very rapid transition, such that $\gamma > 5$ is effectively a two regime-switching model. γ in the range 0–5 exhibits smoother behaviour in the transition function, such that information contained in the transition function is useful in itself as opposed to just containing information regarding the timing of the switch between the two regimes. Thus, placing bounds on $\widehat{\gamma}$ may prove useful for estimating LSTR models.

The model can be approximated by replacing the logistic transition function with a third order Taylor expansion:[8]

$$y_t \simeq \boldsymbol{\beta}'\mathbf{X}_t + \left(\boldsymbol{\theta}'\mathbf{X}_t\right)\left[\frac{1}{2} + \frac{z_t}{4} - \frac{z_t^3}{48}\right] + v_t, \quad v_t \sim \mathsf{IN}\left[0, \sigma_v^2\right] \tag{3.84}$$

where:

$$z_t = \gamma\left(\frac{s_t - c}{\widehat{\sigma}_s}\right). \tag{3.85}$$

[8]Observe that $G''(z)|_{z=0} = 0$, where $G(z) = \left[1 + e^{-z}\right]^{-1}$, and so the z_t^2 term drops out of the Taylor expansion. There is still a quadratic component in s_t as the cubic expansion of $z_t^3 = \gamma^3\left(\frac{s_t - c}{\widehat{\sigma}_s}\right)^3$ is included. $G'(z)|_{z=0} = \frac{1}{4}$ and $G'''(z)|_{z=0} = -\frac{1}{8}$. Furthermore, the fourth order term is $G''''(z)|_{z=0} = 0$, so the next relevant term is the fifth order term. $G'''''(z)|_{z=0} = \frac{1}{4}$, and so inserting into the Taylor expansion results in $G(z) \approx \frac{1}{2} + \frac{z_t}{4} - \frac{z_t^3}{48} + \frac{z_t^5}{480}$. In practice, the orthogonal component of the fifth derivative relative to the third derivative is likely to be very small, and therefore the third order Taylor expansion is sufficient.

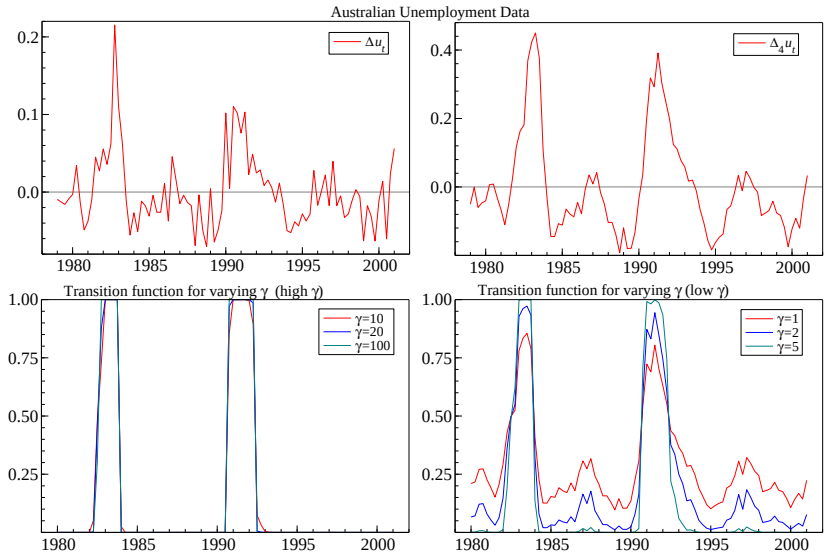

Figure 3.9: Quarterly change in the Australian unemployment rate; annual change in the Australian unemployment rate; LSTR transition functions for high γ; and LSTR transition functions for low γ

This approximation results in a linearised model given by:

$$y_t \simeq \boldsymbol{\alpha}_1' \mathbf{X}_t + \boldsymbol{\alpha}_2' \mathbf{X}_t s_t + \boldsymbol{\alpha}_3' \mathbf{X}_t s_t^2 + \boldsymbol{\alpha}_4' \mathbf{X}_t s_t^3 + v_t, \quad v_t \sim \mathsf{IN}\left[0, \sigma_v^2\right] \tag{3.86}$$

which can be estimated in PcGets.[9] In practice, we would wish to start with a more general approximation.

Selection

The LDGP will consist of two unknown components, the functional form and the relevant variables. The polynomial approximation overcomes the former and *Gets* modelling solves

[9]The transition variable is scaled by $\widehat{\sigma}_s$. When estimating the polynomial approximation, we can pull the scale factor into the coefficient estimates. Assuming $\mathbf{X}_t$ is a scalar for tractability, the mappings from the coefficients in (3.81) to (3.86) are given by:

$$\begin{aligned}
\alpha_1 &= \beta + \frac{\theta}{2} - \frac{\theta\gamma c}{4\widehat{\sigma}_s} + \frac{\theta\gamma^3 c^3}{48\widehat{\sigma}_s^3} \\
\alpha_2 &= \frac{\theta\gamma}{4\widehat{\sigma}_s} - \frac{3\theta\gamma^3 c^2}{48\widehat{\sigma}_s^3} \\
\alpha_3 &= \frac{3\theta\gamma^3 c}{48\widehat{\sigma}_s^3} \\
\alpha_4 &= -\frac{\theta\gamma^3}{48\widehat{\sigma}_s^3}.
\end{aligned}$$

Thus, there is a loss of efficiency by not imposing the cross-parameter restrictions, but that is likely to be very small, and the dominant error is the difference between the LSTR term and its approximation when the LSTR is the LDGP. However, (3.86) remains an identified approximation if LSTR is the incorrect functional form, but a nonlinear model is appropriate.

the latter. We need to formulate a GUM that will include all potentially relevant variables and transition variables for all possible lag lengths. The general linearised GUM based on the LSTR(1) model, for a set of n potential regressors $\mathbf{W}_t$ (including an intercept), and m potential transition variables $\mathbf{S}_t$, is given by:

$$y_t = \sum_{i=1}^{n} \alpha_i W_{i,t} + \sum_{i=1}^{n}\sum_{j=1}^{m} \delta_{ij} W_{i,t} S_{j,t} + \sum_{i=1}^{n}\sum_{j=1}^{m} \lambda_{ij} W_{i,t} S_{j,t}^2 + \sum_{i=1}^{n}\sum_{j=1}^{m} \phi_{ij} W_{i,t} S_{j,t}^3 + \epsilon_t, \quad (3.87)$$

where $\epsilon_t \sim \mathsf{IN}\left[0, \sigma_\epsilon^2\right]$. There are $n + 3nm$ variables in the GUM. The set of potential regressors will include variables that enter in either the linear function or the nonlinear multiplicative function, or both. For variables that enter both components (including a possible intercept), the Taylor approximation will give a parameter estimate that combines both components.

A direct test of linearity against the LSTR model is whether the PcGets selection results in the nonlinear functions being retained:

$$H_0 : \boldsymbol{\delta} = \boldsymbol{\lambda} = \boldsymbol{\phi} = \mathbf{0}. \quad (3.88)$$

The model selected by PcGets should capture the nonlinearity inherent in an LSTR model while enabling a much more general specification to be tested. The approximation to the LSTR model can be tested by estimating the corresponding LSTR model, and then augmenting the final model selected by PcGets with the nonlinear component of the LSTR model. Suppose k relevant variables were retained ($k \leq n$) and one transition variable was selected, given by s_1, then the test of the approximation to the LSTR model would be:

$$H_0 : \boldsymbol{\kappa} = \boldsymbol{\mu} = \boldsymbol{\psi} = \mathbf{0}, \quad (3.89)$$

for the regression:

$$\begin{aligned} y_t = \quad & \sum_{i=1}^{k} \tau_i W_{i,t} + \sum_{i=1}^{k} \kappa_i W_{i,t} s_{1,t} + \sum_{i=1}^{k} \mu_i W_{i,t} s_{1,t}^2 + \sum_{i=1}^{k} \psi_i W_{i,t} s_{1,t}^3 \\ & + \sum_{i=1}^{k} \left(\widetilde{\theta}_i W_{i,t}\right) \left[1 + \exp\left\{-\widetilde{\gamma}\left(\frac{s_{1,t} - \widetilde{c}}{\widehat{\sigma}_{s_1}}\right)\right\}\right]^{-1} + \eta_t. \end{aligned} \quad (3.90)$$

where $\widetilde{\ }$ denotes the estimated parameters from the estimated LSTR model. Thus,

our proposed procedure allows a general test for nonlinearity without the attendant identification problems of the Granger and Teräsvirta (1993) approach, yet allows one to conclude with an LSTR model if that is the best representation.

Monte Carlo results

To examine the ability of the Taylor expansion to approximate an LSTR model, we undertake a simple Monte Carlo experiment based on Granger and Teräsvirta (1993, ch.7). The DGP is given by:

$$y_t = 1 + 2x_t + x_{t-1} + 0.5x_{t-1} - (2x_t + x_{t-1} + 0.5x_{t-2}) \left[1 + \exp\left\{-4\left(x_{t-1} + 3\right)\right\}\right]^{-1} + u_t, \tag{3.91}$$

where x_t is generated as a stationary AR(1) process:

$$x_t = \alpha x_{t-1} + v_t, \qquad v_t \sim \mathsf{IN}\left[0, \sigma_v^2\right]. \tag{3.92}$$

$\alpha = 0, 0.8$ and $\mathsf{V}\left[x_t\right] = 2.78$. Hence, $\sigma_v^2 = 1$ if $\alpha = 0.8$. $u_t \sim \mathsf{IN}\left[0, \sigma_u^2\right]$, in which we set $\sigma_u^2 = 0.0625$, 0.25, 1, and $\mathsf{cov}\left[u_t, v_s\right] = 0\ \forall t, s$. $T = 100$ and the first 100 generated observations are discarded to avoid initialisation effects. $M = 1,000$ replications are undertaken. The models estimated include the LSTR model (3.93), the linear model derived by setting the transition function equal to zero (3.94), and the polynomial approximation in (3.95). Furthermore, selection using the PcGets liberal and conservative strategies is conducted on (3.95).

$$\begin{aligned} y_t &= \beta_0 + \beta_1 x_t + \beta_2 x_{t-1} + \beta_3 x_{t-2} \\ &\quad - \left(\beta_4 x_t + \beta_5 x_{t-1} + \beta_6 x_{t-2}\right) \left[1 + \exp\left\{-\gamma\left(x_{t-1} - c\right)\right\}\right]^{-1} + \epsilon_t \end{aligned} \tag{3.93}$$

$$y_t = \delta_0 + \delta_1 x_t + \delta_2 x_{t-1} + \delta_3 x_{t-2} + \eta_t \tag{3.94}$$

$$\begin{aligned} y_t &= \phi_0 + \phi_1 x_t + \phi_2 x_{t-1} + \phi_3 x_{t-2} + \phi_4 x_t x_{t-1} + \phi_5 x_t x_{t-1}^2 + \phi_6 x_t x_{t-1}^3 + \phi_7 x_{t-1}^2 \\ &\quad + \phi_8 x_{t-1}^3 + \phi_9 x_{t-1}^4 + \phi_{10} x_{t-2} x_{t-1} + \phi_{11} x_{t-2} x_{t-1}^2 + \phi_{12} x_{t-2} x_{t-1}^3 + \varepsilon_t. \end{aligned} \tag{3.95}$$

Table 3.3 records the equation standard errors for the models outlined above. The Monte Carlo LSTR model is an excellent fit, with the equation standard error equal to the true error in all cases. This is because the functional form of (3.93) is known in the

Table 3.3: Equation standard errors for models approximating an LSTR(1) DGP

σ_u	*LSTR(1)*	*Linear*	*Polynomial*	*Liberal*	*Conservative*
$\alpha = \mathbf{0.8}$					
0.25	0.251	1.261	0.414	0.451	0.457
	(0.018)	(0.784)	(0.186)	(0.204)	(0.209)
0.5	0.506	1.365	0.609	0.625	0.634
	(0.067)	(0.725)	(0.147)	(0.209)	(0.217)
1	0.995	1.660	1.060	1.066	1.083
	(0.074)	(0.613)	(0.120)	(0.267)	(0.277)
$\alpha = \mathbf{0}$					
0.25	0.250	0.734	0.310	0.318	0.322
	(0.018)	(0.281)	(0.053)	(0.037)	(0.039)
0.5	0.501	0.864	0.533	0.536	0.543
	(0.036)	(0.245)	(0.050)	(0.054)	(0.056)
1	1.01	1.234	1.015	1.015	1.033
	(0.084)	(0.194)	(0.079)	(0.158)	(0.166)

Notes: standard deviations of the errors are reported in parentheses.

experiments and so the only error comes through estimation uncertainty. In practice, it is unlikely that the exact specification of the LDGP is known, but it provides the optimal infeasible baseline.

The linear model is a poor approximation in all cases. The polynomial approximation performs extremely well when σ_u^2 is large, with an equation standard error just 6% and 2% larger than the true DGP for the $\alpha = 0.8$ and $\alpha = 0$ cases respectively. The approximation is much poorer when the error variance is small, with an equation standard error that is more than 60% larger than the true DGP for $\sigma_u^2 = 0.0625$ and $\alpha = 0.8$. The residual is a composite of the squared approximation error and the DGP shock, so as the latter falls, the former dominates. The squared approximation error is approximately 0.1 for $\alpha = 0.8$, and 0.04 for $\alpha = 0$. This would be large empirically in a log model, although is dependent on the scaling of σ. Undertaking selection on the polynomial LDGP slightly increases the equation standard error but it delivers a more parsimonious model. There is very little cost to selection and the non-deletion probabilities for PcGets are close to the theoretical upper bounds. A further extension would be to see how well PcGets performs when commencing from a more general model. The ability of the polynomial class of functions to approximate nonlinear models will clearly depend on the specific LDGP that is modelled. However, for LDGPs in which there is more uncertainty, the polynomial model performs extremely well.

3.4 Problems of selecting nonlinear models

There are a number of problems that arise when selecting from a GUM that consists of a large set of polynomial regressors. These problems include collinearity, non-normality and excess retention of irrelevant regressors. Solutions to all of these problems are proposed below, confirming the feasibility of the nonlinear model selection strategy.

3.4.1 Collinearity

Multicollinearity was first outlined by Frisch (1934) within the context of static general equilibrium linear relations. Confluence analysis was developed to address the problem, although this method is not in common practice now (see Hendry and Morgan, 1989). The definition of collinearity has shifted over the years, and we can define perfect collinearity as $|\mathbf{X}'\mathbf{X}| = 0$, and perfect orthogonality as a diagonal $(\mathbf{X}'\mathbf{X})$ matrix. Since collinearity is not invariant under linear transformations, it is difficult to identify the degree of collinearity. A linear model is invariant under linear transformations, and so a model could be defined by various isomorphic representations, which nevertheless deliver very different inter-correlations. Hence, collinearity is a property of the parameterisation of the model, and not the variables *per se*.

Nonlinear transformations can generate substantial collinearity between the linear and nonlinear functions. We initially consider a simple case of a transformation given by $f(x) = x^2$. This polynomial transform is common in economics; e.g., age and the square of age often enter in labour force participation models. The degree of collinearity varies as the statistical properties of the process vary. For example, the collinearity between x and x^2 is 0 when $\mathsf{E}[x] = 0$, but increases to near perfect collinearity as $\mathsf{E}[x]$ increases. Both analytic and Monte Carlo results are assessed.

3.4.2 Analytical results for the correlation of x and x^2

We formulate the DGP as the linear conditional relation:

$$y_i = x_i + e_i = 0 + x_i + 0x_i^2 + e_i \tag{3.96}$$

with:

$$x_i \sim \mathsf{IN}\left[0, \sigma_x^2\right] \tag{3.97}$$

$$e_i \sim \mathsf{IN}\left[0, \sigma_e^2\right] \tag{3.98}$$

$$\mathsf{E}\left[x_i, e_i\right] = 0, \quad \forall i. \tag{3.99}$$

Since (3.96) is invariant under linear transformations, it can also be written for $z_i = x_i + \mu$:

$$\begin{aligned} y_i &= -\mu + (x_i + \mu) + 0\,(x_i + \mu)^2 + e_i \\ &= -\overline{z} + z_i + 0 z_i^2 + e_i \\ &= 0 + (z_i - \overline{z}) + 0\,(z_i - \overline{z})^2 + e_i. \end{aligned} \tag{3.100}$$

We consider three models, including the zero-mean case:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + u_i, \tag{3.101}$$

the complete zero-mean case:

$$y_i = \lambda_0 + \lambda_1 x_i + \lambda_2 \left(x_i^2 - \overline{x^2}\right) + u_i = \lambda_0 + \lambda_1 x_i + \lambda_2 w_i + u_i, \tag{3.102}$$

and the non-zero mean case:

$$y_i = \gamma_0 + \gamma_1 z_i + \gamma_2 z_i^2 + u_i. \tag{3.103}$$

First, examining the general case (3.103), in which there is a non-zero mean:

$$\begin{aligned} \mathsf{E}\left[T^{-1}\mathbf{X}'\mathbf{X}_{(\mu)}\right] &= \mathsf{E}\left[\begin{pmatrix} 1.0 & \overline{z} & \overline{z^2} \\ \overline{z} & T^{-1}\sum z_i^2 & T^{-1}\sum z_i^3 \\ \overline{z^2} & T^{-1}\sum z_i^3 & T^{-1}\sum z_i^4 \end{pmatrix}\right] \\ &= \begin{bmatrix} 1.0 & \mu & \mu^2 + \sigma_x^2 \\ \mu & \mu^2 + \sigma_x^2 & \mu^3 + 3\mu\sigma_x^2 \\ \mu^2 + \sigma_x^2 & \mu^3 + 3\mu\sigma_x^2 & 3\sigma_x^4 + \mu^4 + 6\mu^2\sigma_x^2 \end{bmatrix} \end{aligned} \tag{3.104}$$

with the inverse computed as:

$$\left(\mathsf{E}\left[T^{-1}\mathbf{X}'\mathbf{X}_{(\mu)}\right]\right)^{-1} = \frac{1}{2\sigma_x^6}\begin{bmatrix} \mu^4\sigma_x^2 + 3\sigma_x^6 & -2\mu^3\sigma_x^2 & \mu^2\sigma_x^2 - \sigma_x^4 \\ -2\mu^3\sigma_x^2 & 2\sigma_x^4 + 4\mu^2\sigma_x^2 & -2\mu\sigma_x^2 \\ \mu^2\sigma_x^2 - \sigma_x^4 & -2\mu\sigma_x^2 & \sigma_x^2 \end{bmatrix}. \tag{3.105}$$

There is substantial collinearity between the variables, except for the squared term which

is irrelevant in the DGP.

Next, consider the zero-mean model in equation (3.101):

$$\mathsf{E}\left[T^{-1}\mathbf{X}'\mathbf{X}_{(0)}\right] = \mathsf{E}\left[\begin{pmatrix} 1.0 & \overline{x} & \overline{x^2} \\ \overline{x} & T^{-1}\sum x_i^2 & T^{-1}\sum x_i^3 \\ \overline{x^2} & T^{-1}\sum x_i^3 & T^{-1}\sum x_i^4 \end{pmatrix}\right] = \begin{bmatrix} 1.0 & 0.0 & \sigma_x^2 \\ 0.0 & \sigma_x^2 & 0.0 \\ \sigma_x^2 & 0.0 & 3\sigma_x^4 \end{bmatrix} \tag{3.106}$$

so the inverse is:

$$\left(\mathsf{E}\left[T^{-1}\mathbf{X}'\mathbf{X}_{(0)}\right]\right)^{-1} = \frac{1}{2\sigma_x^6}\begin{bmatrix} 3\sigma_x^6 & 0 & -\sigma_x^4 \\ 0 & 2\sigma_x^4 & 0 \\ -\sigma_x^4 & 0 & \sigma_x^2 \end{bmatrix}. \tag{3.107}$$

There is no collinearity between x and x^2. There is an effect on the intercept, as the 'correlation' between the intercept and x^2 is -0.577, but this does not cause a problem for the PcGets selection algorithm.

Then, examining the complete zero-mean model in equation (3.102):

$$\begin{aligned} \mathsf{E}\left[T^{-1}\mathbf{X}'\mathbf{X}_{(0,0)}\right] &= \mathsf{E}\left[\begin{pmatrix} 1.0 & \overline{x} & \overline{w} \\ \overline{x} & T^{-1}\sum x_i^2 & T^{-1}\sum x_i w_i \\ \overline{w} & T^{-1}\sum x_i w_i & T^{-1}\sum w_i^2 \end{pmatrix}\right] \\ &= \begin{bmatrix} 1.0 & 0.0 & 0.0 \\ 0.0 & \sigma_x^2 & 0.0 \\ 0.0 & 0.0 & 3\sigma_x^4 \end{bmatrix} \end{aligned} \tag{3.108}$$

as:

$$T^{-1}\sum x_i w_i = T^{-1}\sum x_i\left(x_i^2 - \overline{x^2}\right) = T^{-1}\sum x_i^3 - \overline{x^2}T^{-1}\sum x_i = T^{-1}\sum x_i^3 - \overline{x}\overline{x^2},$$

so the inverse is:

$$\left(\mathsf{E}\left[T^{-1}\mathbf{X}'\mathbf{X}_{(0,0)}\right]\right)^{-1} = \frac{1}{3\sigma_x^6}\begin{bmatrix} 3\sigma_x^6 & 0.0 & 0.0 \\ 0.0 & 3\sigma_x^4 & 0.0 \\ 0.0 & 0.0 & \sigma_x^2 \end{bmatrix}. \tag{3.109}$$

We noted that collinearity is a property of the parameterisation of the model, and so we seek a near orthogonal representation of the general model. This can be achieved simply by taking deviations from means, which re-creates the specification in terms of the original variables x and x^2. Observe that $z_i = x_i + \mu$ where $\overline{z} = \mu$ and $\overline{z^2} = \mu^2 + \sigma_x^2$. Hence, we can calculate:

$$\mathsf{E}\left[T^{-1}\mathbf{X}'\mathbf{X}_{(\overline{\mu})}\right] = \mathsf{E}\left[\begin{pmatrix} 1.0 & z_i - \overline{z} & z_i^2 - \overline{z^2} \\ z_i - \overline{z} & T^{-1}\sum (z_i - \overline{z})^2 & T^{-1}\sum (z_i - \overline{z})\left(z_i^2 - \overline{z^2}\right) \\ z_i^2 - \overline{z^2} & T^{-1}\sum (z_i - \overline{z})\left(z_i^2 - \overline{z^2}\right) & T^{-1}\sum \left(z_i^2 - \overline{z^2}\right)^2 \end{pmatrix}\right]$$

$$= \begin{bmatrix} 1.0 & 0.0 & 0.0 \\ 0.0 & \sigma_x^2 & 2\mu\sigma_x^2 \\ 0.0 & 2\mu\sigma_x^2 & 3\sigma_x^6 - \sigma_x^4 + 4\mu^2\sigma_x^2 \end{bmatrix} \tag{3.110}$$

with the inverse:

$$\left(\mathsf{E}\left[T^{-1}\mathbf{X}'\mathbf{X}_{(\overline{\mu})}\right]\right)^{-1} = \frac{1}{\sigma_x^6\left(3\sigma_x^2 - 1\right)} \begin{bmatrix} 3\sigma_x^8 - \sigma_x^6 & 0.0 & 0.0 \\ 0.0 & 3\sigma_x^6 - \sigma_x^4 + 4\mu^2\sigma_x^2 & -2\mu\sigma_x^2 \\ 0.0 & -2\mu\sigma_x^2 & \sigma_x^2 \end{bmatrix}. \tag{3.111}$$

Taking deviations from means delivers some reduction in collinearity, which is particularly marked for the intercept, but worse for the linear term $(z_i - \overline{z})$, demonstrating that 'single' de-meaning is not sufficient to obtain a near orthogonal representation. Again the irrelevant squared term 'benefits'.

Figure 3.10a records the correlation coefficients between the intercept, z and z^2 for the general case, with $\sigma_x^2 = 1$ and varying μ. Figure 3.10b records the correlation coefficients with $\mu = 1$ and σ_x varying.[10] The degree of collinearity increases sharply as μ increases: e.g., when $\mu = 10$ and $\sigma_x = 1$, the correlation between z and z^2 is -0.9775, the 'correlation' of z with the intercept is -0.99735 and the 'correlation' of z^2 with the intercept is 0.98985. Hence, the general case can generate near perfect collinearity. Taking deviations from means, as in (3.110), results in a correlation of 0.99875 for z and z^2, and so this correlation is higher after undertaking an 'orthogonalising' transform with the intercept. To remove the collinearity between z and z^2, we also need to de-mean z^2. Hence, the linear term remains $(z_i - \overline{z})$, but the squared term becomes $(z_i - \overline{z})^2 - [\mathsf{E}(z_i - \overline{z})]^2$, which will result in a model that is identical to (3.102). Double de-meaning has removed the collinearity generated by the non-zero mean in the white-noise process.

3.4.3 Monte Carlo evidence on collinearity

We next examine Monte Carlo evidence for PcGets when selecting from a GUM containing highly collinear nonlinear variables. We consider three models in which the DGP is a static process, a stationary dynamic process, and a unit-root process, and focus on quadratic functions to match the analytics. Results for both the liberal and conservative strategies are reported.

[10] The absolute values of the correlations are recorded.

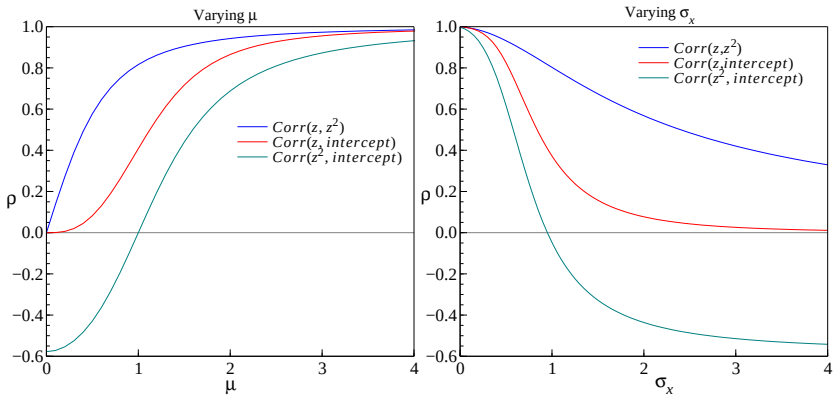

Figure 3.10: Correlations between the intercept, z and z^2 for the general model. Varying μ with $\sigma_x^2 = 1$, and varying σ_x with $\mu = 1$

A static model

The DGP is given by:

$$y_t = \beta x_t + u_t, \qquad u_t \sim \mathsf{IN}\left[0, \sigma_u^2\right] \tag{3.112}$$

$$x_t = \mu + \nu_t, \qquad \nu_t \sim \mathsf{IN}\left[0, \sigma_v^2\right] \tag{3.113}$$

for $t = 1, ..., T$. We set $\sigma_u^2 = \sigma_v^2 = 1$ and $\mu = 0$ or 10. Two sample sizes are assessed, $T = 100$ and $1,000$. The number of replications, M, is 10,000. To ensure the probability of retaining the relevant variable is near unity we set $\mathsf{E}\left[|\mathsf{t}_\beta|\right] = \psi_\beta = 5$, which corresponds to $\beta = 0.5$ for $T = 100$ and $\beta = 0.158$ for $T = 1,000$ in the orthogonal case. The theoretical maximum power is 0.9987 at the 5% significance level and 0.9912 at the 1% significance level. The GUM contains three variables, two of which are nuisance, and is given by:

$$y_t = \alpha_0 + \alpha_1 x_t + \alpha_2 x_t^2 + \epsilon_t, \quad \epsilon_t \sim \mathsf{IN}\left[0, \sigma_\epsilon^2\right]. \tag{3.114}$$

To overcome the collinearity problem induced by the non-zero mean, we seek a near orthogonal representation of the model. As $\mathsf{E}\left[x_t\right] = \mu$ and $\mathsf{E}\left[x_t^2\right] = \mu^2 + \sigma_v^2$, we would ideally de-mean using the rules:

$$\widetilde{x_t} = x_t - \mu, \quad \text{and} \quad \widetilde{x_t^2} = (x_t - \mu)^2 - \sigma_v^2. \tag{3.115}$$

However, in practice population values are unknown and sample estimates are used in-

stead:

$$\widetilde{x} = x_t - \widehat{\mu}, \quad \text{and} \quad \widetilde{x^2} = (x_t - \widehat{\mu})^2 - \widehat{\sigma}_v^2. \tag{3.116}$$

Results Figure 3.11 records the probability of retaining variables for $\mu = 0$, $\mu = 10$, and the de-meaned case using both rules (3.115) and (3.116). For $\mu = 0$, the probability of retaining x_t is almost unity and is higher for the liberal strategy than the conservative strategy. The probability of retaining the irrelevant variables is marginally higher than 5% and 1%, but it is essentially controlled at the selected significance levels, and reliability weighting would reduce this size further. However, for $\mu = 10$, there is a dramatic fall in the probability of retaining x_t to below 60%, and a corresponding increase in the probability of retaining the intercept and x_t^2 to over 40%. The conservative strategy can deliver higher power than the liberal strategy (for the $T = 100$ case), implying that it deletes the irrelevant collinear variable more often and so finds the DGP variable slightly more often. The transformation to an orthogonal representation results in correct retention probabilities for the liberal and conservative strategies, matching the $\mu = 0$ results. Furthermore, using sample moments as opposed to population moments does not impact on the retention probabilities. Hence, by undertaking these simple operational rules, the adverse selection properties resulting from the collinearity between x_t and x_t^2 are mitigated.

Varying non-centralities for the DGP variables

We next consider the properties of PcGets when the non-centralities, ψ, of the DGP variables vary. The DGP is formulated in (3.117), in which two linear variables enter the DGP:

$$y_t = \beta_1 x_{1,t} + \beta_2 x_{2,t} + u_t, \qquad u_t \sim \mathsf{IN}\left[0, \sigma_u^2\right] \tag{3.117}$$

$$\mathbf{x}_t = \boldsymbol{\mu} + \boldsymbol{\nu}_t, \qquad \boldsymbol{\nu}_t \sim \mathsf{IN}_2\left[\mathbf{0}, \boldsymbol{\Omega}_\nu\right] \tag{3.118}$$

for $t = 1, ..., T$, where $\boldsymbol{\Omega}_{\nu,ij} = 0$ for $i \neq j$ and $\boldsymbol{\Omega}_{\nu,ii} = \sigma_u^2 = 1$ for $i = 1, 2$. The GUM contains three nuisance parameters and is given by:

$$y_t = \alpha_0 + \alpha_1 x_{1,t} + \alpha_2 x_{2,t} + \alpha_3 x_{1,t}^2 + \alpha_4 x_{2,t}^2 + \epsilon_t, \quad \epsilon_t \sim \mathsf{IN}\left[0, \sigma_\epsilon^2\right]. \tag{3.119}$$

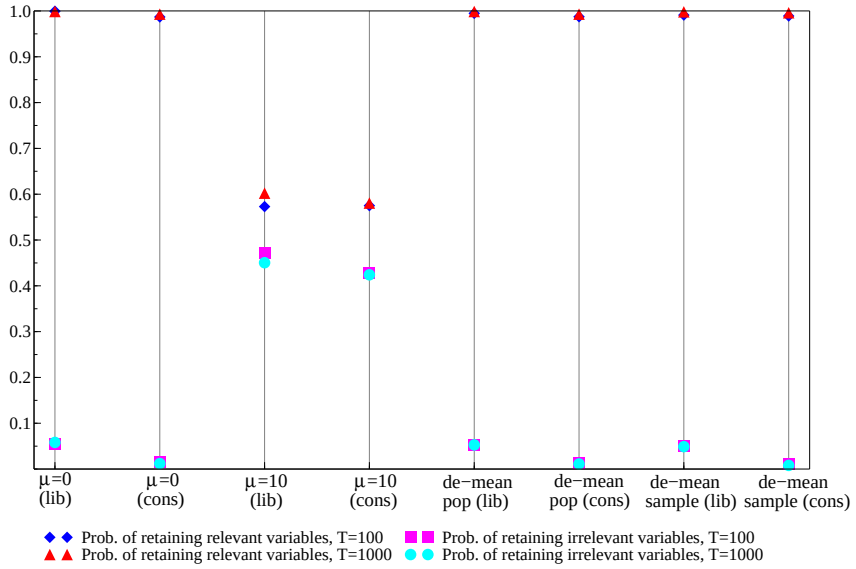

Figure 3.11: The probability of PcGets retaining variables after selection for the static nonlinear case, comparing orthogonal and collinear models for both the liberal and conservative strategies

Four cases are considered: defining $\mathsf{E}[|t_{\beta_i}|] = \psi_{\beta_i}$, under orthogonality the cases include (i) $\psi_{\beta_1} = \psi_{\beta_2} = 2$; (ii) $\psi_{\beta_1} = \psi_{\beta_2} = 3$; (iii) $\psi_{\beta_1} = \psi_{\beta_2} = 4$; and (iv) $\psi_{\beta_1} = 3, \quad \psi_{\beta_2} = 6$. Results are reported for T=100 and M =10,000 replications.[11]

Results Figure 3.12 records the retention probabilities of the liberal and conservative strategies for the four cases outlined, when $\mu = 10$ and after de-meaning using rule (3.116). Transforming to an orthogonal representation increases the probability of retaining the relevant variables (other than for the conservative strategy with $\psi_{\beta_i} = 2$), and tightens the probability of retaining the irrelevant variables to 5% and 1% for the liberal and conservative strategies respectively. Figure 3.13 records the power for a t-test of a single null hypothesis, H_0, where $\psi_{\beta_i} = 0$ under the null, using a 2-sided test at critical value c_α. To calculate the power to reject the null when $\psi_{\beta_i} > 0$, we use:

$$\mathsf{P}(\mathsf{t}_{\beta_i} \geq c_\alpha \mid \mathsf{E}[|\mathsf{t}_{\beta_i}|] = \psi_{\beta_i}) \simeq \mathsf{P}(\mathsf{t}_{\beta_i} - \psi_{\beta_i} \geq c_\alpha - \psi_{\beta_i} \mid \mathsf{H}_0). \tag{3.120}$$

There is a 50% chance of retaining a single variable with $|\mathsf{t}| = 2$ when $\alpha = 0.05$, but this falls to 27% when $\alpha = 0.01$. The power to detect relevant variables increases with the non-centrality, and $|\mathsf{t}|$s of 4 are retained more than 90% of the time. We also consider

[11] Results for $T = 1,000$ are analogous to those for $T = 100$, and are available on request.

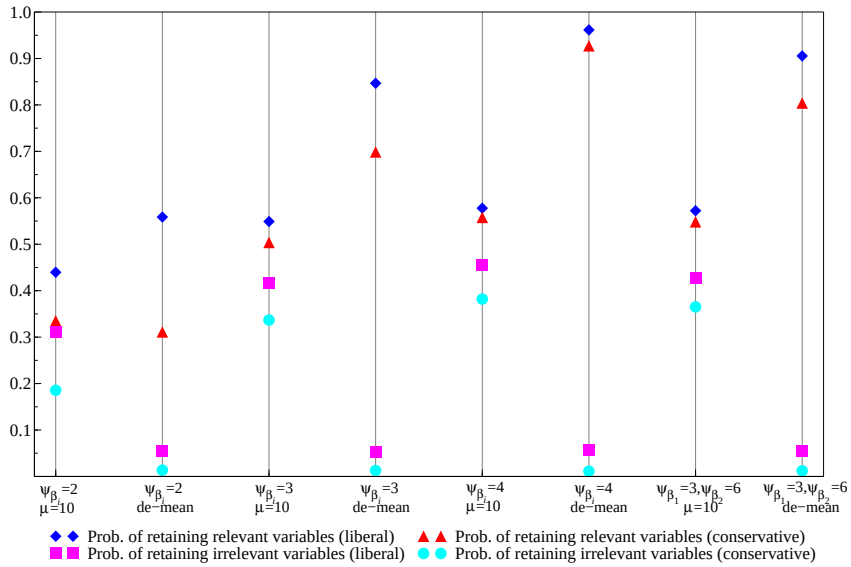

Figure 3.12: The probability of retaining relevant and irrelevant variables for the static nonlinear case with two exogenous variables

the theoretical probability of retaining two relevant variables, matching the four cases examined. PcGets retention rates for the de-meaned variables are comparable to the theoretical single t-test results, matching the linear studies by Hendry and Krolzig (1999, 2003b).

A stationary AR(1) process

Consider a stationary AR(1) process for the regressor, in which the DGP is given by:

$$y_t = \beta x_t + u_t, \qquad u_t \sim \mathsf{IN}\left[0, \sigma_u^2\right] \tag{3.121}$$

$$x_t = \mu + \rho x_{t-1} + \nu_t, \qquad \nu_t \sim \mathsf{IN}\left[0, \sigma_v^2\right] \tag{3.122}$$

for $t = 1, ..., T$. We set $\sigma_u^2 = \sigma_v^2 = 1$ and $\rho = 0.8$. We examine the zero mean case ($\mu = 0$), and the case where $\mathsf{E}[x_t] = \frac{\mu}{1-\rho} = 10$, setting $\mu = 2$. We use two rules to de-mean, the first removes the population means and the second removes the sample means:

$$1) \quad \widetilde{x}_t = x_t - \frac{\mu}{1-\rho}, \quad \text{and} \quad \widetilde{x_t^2} = \left(x_t - \frac{\mu}{1-\rho}\right)^2 - \frac{\sigma_v^2}{1-\rho^2} \tag{3.123}$$

$$2) \quad \widetilde{x}_t = x_t - \frac{\widehat{\mu}}{1-\widehat{\rho}}, \quad \text{and} \quad \widetilde{x_t^2} = \left(x_t - \frac{\widehat{\mu}}{1-\widehat{\rho}}\right)^2 - \frac{\widehat{\sigma}_v^2}{1-\widehat{\rho}^2}. \tag{3.124}$$

The first 50 generated observations are discarded for each replication. We set $\beta = 0.5$ for $T = 100$ and $\beta = 0.158$ for $T = 1,000$, with $M = 10,000$ replications. We consider two

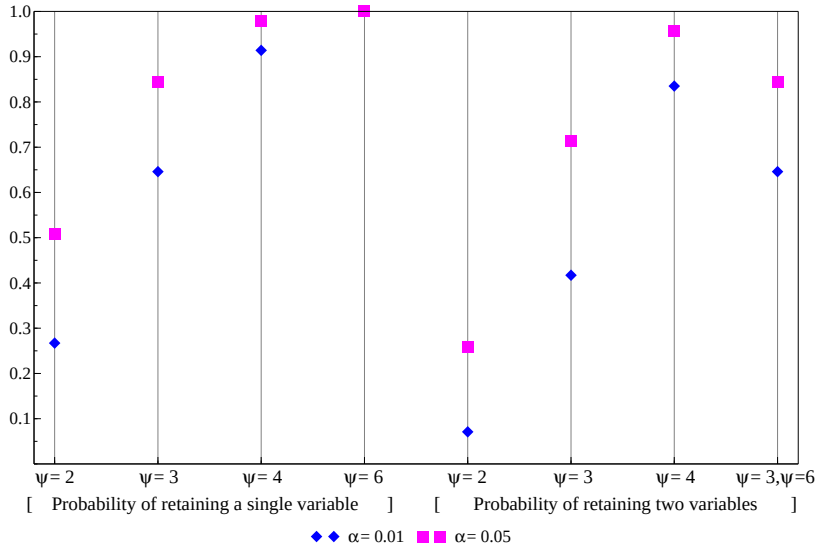

Figure 3.13: t-test powers for a single null hypothesis test and two null hypothesis tests

GUMs, (3.125) in which there are five nuisance parameters and (3.125) with no dynamics, where $\alpha_1 = \alpha_3 = \alpha_5 = 0$.

$$y_t = \alpha_0 + \alpha_1 y_{t-1} + \alpha_2 x_t + \alpha_3 x_{t-1} + \alpha_4 x_t^2 + \alpha_5 x_{t-1}^2 + \epsilon_t, \quad \epsilon_t \sim \mathsf{IN}\left[0, \sigma_\epsilon^2\right]. \tag{3.125}$$

Results Figure 3.14 records the retention probabilities of PcGets for the AR(1) process when $T = 100$.[12] The retention probabilities for the GUM containing three variables and six variables are almost identical, indicating that exogenous dynamics do not affect the selection probabilities to the extent that the nonlinear functions do. With a zero mean process, the probability of retaining x_t is near unity for both strategies, and the probability of retaining irrelevant variables corresponds to 5% and 1% for the liberal and conservative strategies respectively. A non-zero mean results in a decrease in the probability of retaining x_t to below 70%. The probability of retaining irrelevant variables is much higher for the three-variable GUM than the six-variable GUM. The correlation between x_t and x_t^2 is much higher than the correlation between the variables dated t and $t-1$. Therefore, averaging retention probabilities across two variables results in a higher null rejection frequency than averaging across five variables. The collinearity between variables dated t and $t-1$ is significant, with retention probabilities of approximately 14%

[12] Results for $T = 1,000$ are similar and are therefore not reported, but are available on request.

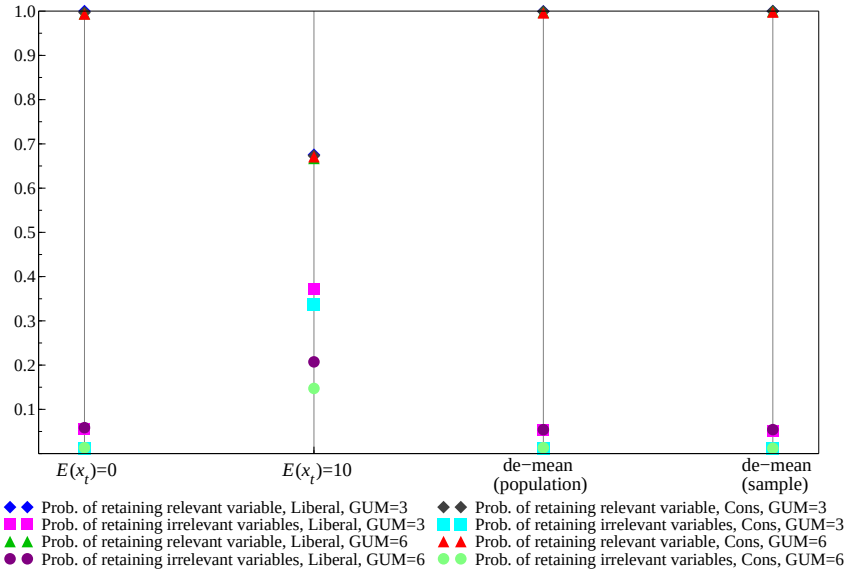

Figure 3.14: The probability of retaining relevant and irrelevant variables for the stationary dynamic nonlinear case, $T = 100$. GUM=6 includes lagged variables and GUM=3 excludes lagged variables

for the liberal strategy and 3% for the conservative strategy, but the average retention probabilities for x_t^2 and intercept are approximately 35% and 33% respectively. The use of sample means has no impact on retention probabilities compared to population values, providing easily implementable, operational rules for orthogonalisation: the last two columns of Figure 3.11 and Figure 3.14 illustrate.

A unit-root process

Finally, we consider a DGP that consists of a unit-root process outlined in (3.127).

$$y_t = \beta x_t + u_t, \qquad u_t \sim \mathsf{IN}\left[0, \sigma_u^2\right] \tag{3.126}$$

$$x_t = x_{t-1} + \nu_t, \qquad \nu_t \sim \mathsf{IN}\left[0, \sigma_v^2\right] \tag{3.127}$$

for $t = 1, ..., T$. We consider two initial conditions, $x_0 = 0$ and 10 and set $\sigma_u^2 = \sigma_v^2 = 1$. $\beta = 0.5$ for $T = 100$, and $\beta = 0.158$ for $T = 1,000$, with $M = 10,000$. The GUM is given by (3.125), and we consider both the full equation, and setting $\alpha_1 = \alpha_3 = \alpha_5 = 0$.

In order to de-mean the data, we remove sample means, outlined in rule (3.128):

$$\widetilde{x}_t = x_t - \frac{1}{T}\sum_{t=1}^{T} x_t, \quad \text{and} \quad \widetilde{x_t^2} = \left(x_t - \frac{1}{T}\sum_{t=1}^{T} x_t\right)^2 - \frac{1}{T}\sum_{t=1}^{T}\left(x_t - \frac{1}{T}\sum_{t=1}^{T} x_t\right)^2. \tag{3.128}$$

Results Figure 3.15 records retention probabilities for the unit-root process for $T = 100$. For an initial condition of zero, the probability of retaining x_t for both the liberal and conservative strategy is unity, and the probability of retaining irrelevant variables is approximately 5% and 1% respectively. Hence, I(1)ness has no impact on selection when there are nonlinear functions. However, imposing an initial condition of 10 dramatically reduces the probability of retaining relevant variables and increases the probability of retaining irrelevant variables, and, for a GUM of six, the probability of retaining x_t is marginally higher for the conservative strategy compared to the liberal strategy. Rule (3.128) results in a retention probability of unity for x_t, with corresponding correct probabilities for the retention of the irrelevant variables. Thus, rule (3.128) ensures near orthogonal nonlinear regressors for I(1) variables. Observe that removing the initial condition would deliver results analogous to those in which there is a zero initial condition.[13] Removing the initial condition would result in a retention probability of approximately 5% and 1% for irrelevant variables and a retention probability near to the theoretical upper bound for relevant variables.

Figure 3.16 records the retention probabilities for the unit root process for $T = 1,000$. The adverse retention probabilities are mitigated as the sample size increases. The random walk has deviated from the initial condition substantially, and while the correlations do still depend on x_0, the impact is rapidly declining with T. As de-meaning using the sample average yields no cost, it would be recommended regardless of the sample size.

A differenced I(1) process

While we have identified the primary difficulty in this aspect of model selection to be collinearity between variables and their corresponding nonlinear transformations, the correlation between variables and their lags is also significant. Retention of x_{t-1} and x_{t-1}^2 is approximately three times higher than the 5% and 1% size of the liberal and conservative strategies respectively, for both the stationary process when $\mathsf{E}[x_t] = 10$ and

[13] In practice the initial condition will be unknown and the first observation, x_1, could be used as an estimate of the initial condition x_0.

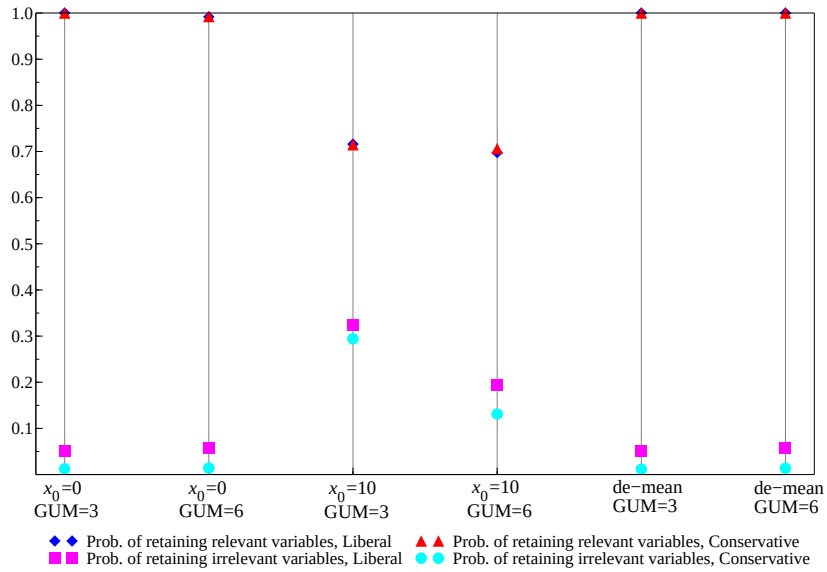

Figure 3.15: Probability of retaining relevant and irrelevant variables for the unit-root case, $T = 100$

the non-stationary process when $x_0 = 10$ for $T = 100$. Hence, we next consider the impact of differencing to remove the collinearity between the variables and their lags.

The DGP is given as:

$$y_t = \beta x_{t-1} + u_t, \qquad u_t \sim \mathsf{IN}\left[0, \sigma_u^2\right] \tag{3.129}$$

for $t = 1, ..., T$, with x_t generated by (3.127). Again, we set $\sigma_u^2 = \sigma_v^2 = 1$, $\beta = 0.5$ for $T = 100$ and $\beta = 0.158$ for $T = 1,000$, with $M = 10,000$. The GUM is given by:

$$y_t = \alpha_0 + \alpha_1 y_{t-1} + \alpha_2 \Delta x_t + \alpha_3 x_{t-1} + \alpha_4 \Delta x_t^2 + \alpha_5 x_{t-1}^2 + \epsilon_t, \quad \epsilon_t \sim \mathsf{IN}\left[0, \sigma_\epsilon^2\right], \tag{3.130}$$

where $\mathsf{cov}\left[\Delta x_t, x_{t-1}\right] = \mathsf{cov}\left[\Delta x_t^2, x_{t-1}^2\right] = 0$. We consider the case in which $x_0 = 10$ and we use rule (3.128) to de-mean the data.

Results Figure 3.17 records the retention probabilities of the unit-root process for $T = 100$. The first column is in levels, where the DGP is (3.129) but the GUM is (3.125). PcGets struggles to identify the DGP variable and the retention probability is substantially larger than 5% and 1% for the irrelevant variables, for the two strategies. The second column uses the GUM in which the lagged variables are orthogonalised, given in (3.130), and the retention probabilities are analogous to the levels model. Thus,

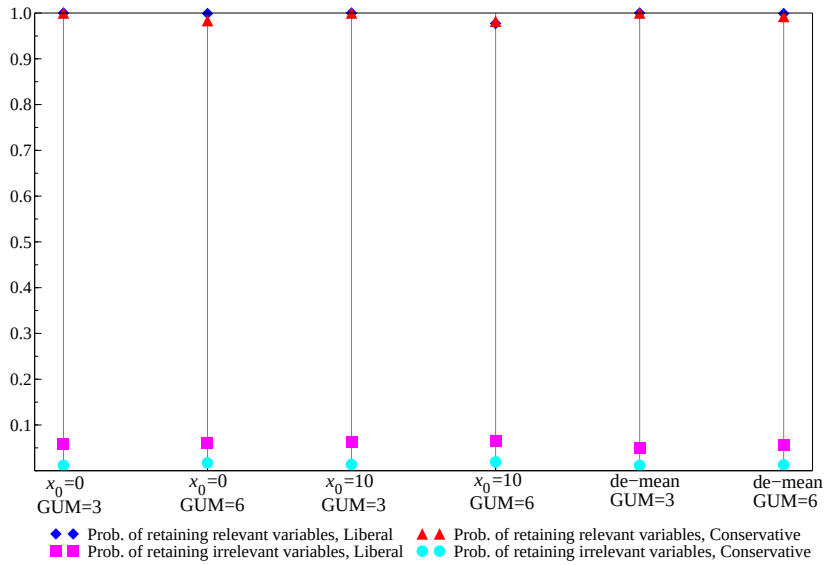

Figure 3.16: Probability of retaining relevant and irrelevant variables for the unit-root case, $T = 1,000$

orthogonalising the lagged variables alone without ensuring orthogonality of the contemporaneous variables yields no improvement in selection. The final column removes sample means from the data and orthogonalises by differencing, and this results in a retention probability of near unity for the relevant variable and a retention probability of approximately 5% and 1% for the irrelevant variables, for the liberal and conservative strategies respectively. Hence, all forms of collinearity need to be removed when undertaking model selection.

Figure 3.18 provides more detail to Figure 3.17, recording the retention probabilities of all GUM variables for the three cases outlined. Retention of the lagged dependent variable has the appropriate probabilities, indicating that LDVs are not problematic for model selection when they do not enter the DGP. In levels, x_t is retained approximately three times too often, at 15% and 4% respectively. Taking differences halves the retention probabilities to 7% and 2%, but de-meaning as well as taking differences reduces the retention probabilities further, to 5% and 1%. The same pattern is evident with x_t^2. The DGP variable is retained about 70% of the time for the levels GUM, with the conservative strategy delivering a higher power than the liberal strategy. Taking differences does not alter this probability, but de-meaning does increase the probability of retaining the DGP variable to unity. Both x_{t-1}^2 and the intercept are retained far too often at about 30%,

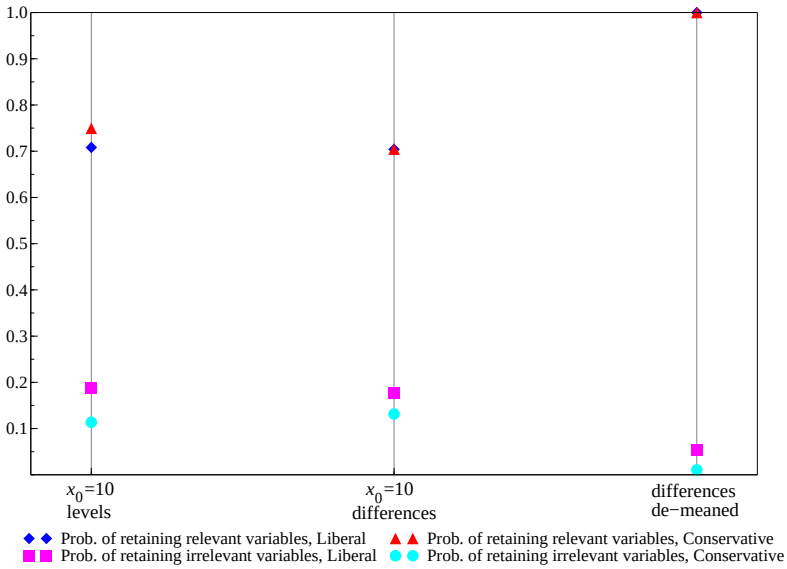

Figure 3.17: Retention probabilities for the relevant and irrelevant variables for the unit root case with a differenced GUM, $T = 100$

which is analogous to the above results. Orthogonalising by differencing does not alter these probabilities as the collinearity between x_{t-1} and x_{t-1}^2 is unaffected, but de-meaning reduces the retention probability to appropriate levels.

The nonlinear selection strategy will automatically double de-mean the generated polynomial functions prior to formulating the GUM. Two caveats apply. First, the orthogonalising rules will not remove all collinearity between higher order polynomials. Orthogonalisation using the Choleski method of solving a system of linear simultaneous algebraic equations (see Rushton, 1951) has been considered, but the simple double de-meaning removes enough collinearity to ensure the PcGets selection has the appropriate properties and so this strategy is implemented. Second, any information contained in the intercepts of the explanatory variables will be removed, although this concern is not fundamental as there is rarely a theory of the intercept postulated when developing econometric models.

3.4.4 Non-normality

Normality is a basic assumption in PcGets, as conventional critical values are used for selection and null rejection frequencies would be incorrect for non-normality. Normality tends to be more vital for selection (when many decisions are made) than inference. A test

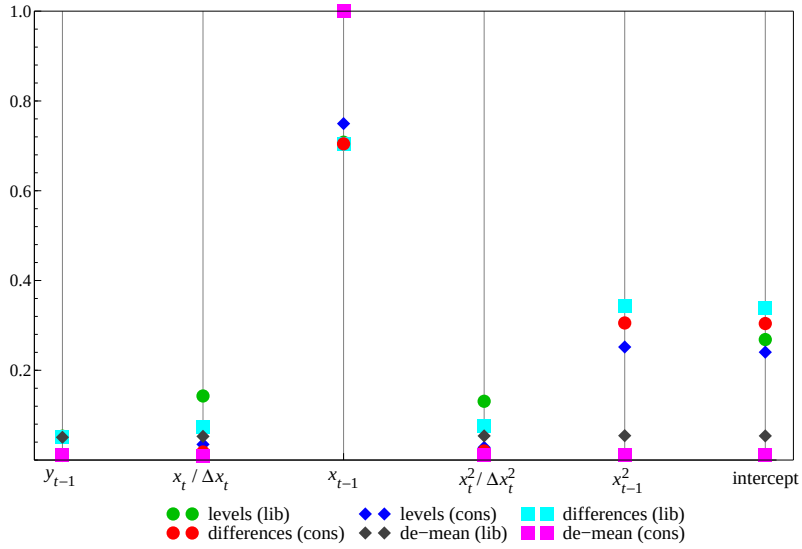

Figure 3.18: Retention probabilities of the GUM variables for the unit root case, $T = 100$

for normality based on Doornik and Hansen (1994) is performed on the GUM, and the diagnostics are checked at every subsequent reduction stage. If a reduction brings about a rejection of a diagnostic test, the search is terminated at the preceeding level. When we consider nonlinear models, normality becomes an essential requirement. Problems arise when extreme observations result in fat-tailed distributions, as there is an increased probability that nonlinear functions will align with extreme observations, effectively acting as indicators and therefore being retained too often. This can be demonstrated by considering a simple case (of relevance below) in which an outlier is modelled as an indicator variable, $I_{\{t=s\}}$, which takes the value 1 in period s and 0 otherwise. Consider a regression between two unconnected variables:

$$y_t = \beta x_t + \delta I_{\{t=s\}} + u_t \tag{3.131}$$

$$x_t = \gamma I_{\{t=s\}} + v_t, \tag{3.132}$$

where $\beta = 0$. If the indicator is omitted from the model, so $y_t = \beta x_t + u_t$, we can calculate the coefficient $\widehat{\beta}$ as:

$$\begin{aligned} \widehat{\beta} = \frac{\sum x_t y_t}{\sum x_t^2} &= \frac{\delta\gamma \sum I_{\{t=s\}}^2 + \sum (\delta v_t + \gamma u_t) I_{\{t=s\}} + \sum v_t u_t}{\gamma^2 \sum I_{\{t=s\}}^2 + 2\gamma \sum v_t I_{\{t=s\}} + \sum v_t^2} \\ &= \frac{\delta\gamma + (\delta v_s + \gamma u_s) + \sum v_t u_t}{\gamma^2 + 2\gamma v_s + \sum v_t^2} \end{aligned} \tag{3.133}$$

as $\sum I^2_{\{t=s\}} = 1$. Also:

$$\mathsf{V}\left[\widehat{\beta}\right] = \frac{\widehat{\sigma}^2_u}{\sum x^2_t} \tag{3.134}$$

and:

$$\mathsf{t}_{\widehat{\beta}} = \frac{\widehat{\beta}\sqrt{\sum x^2_t}}{\widehat{\sigma}_u} = \frac{\sum x_t y_t}{\widehat{\sigma}_u \sqrt{\sum x^2_t}}. \tag{3.135}$$

Hence, if we approximate by $v_s = u_s = 0$, $\widehat{\sigma}_u = 1$, $\frac{\gamma}{\widehat{\sigma}_v} = 1$ and $\sum v_t u_t \simeq 0$ we can calculate:

$$\mathsf{t}^2_{\widehat{\beta}} = \frac{\delta^2 \gamma^2}{\gamma^2 + T}. \tag{3.136}$$

To illustrate this phenomenon, suppose $\delta = 6$, $\gamma = 5$, and $T = 100$. Then:

$$\mathsf{t}^2_{\widehat{\beta}} = \frac{6^2 \times 5^2}{5^2 + 100} = 7.2. \tag{3.137}$$

Thus, outliers need to be quite large for this effect. This is plausible when considering nonlinear transformations. For example, in one draw of an $\mathsf{IN}[0,1]$ process with $T = 100$, the standard deviation of the inverse transformation is $\sigma = 25.608$ and the largest outlier is -169.2, which RETINA could well confront.

Extreme observations: Monte Carlo example

Monte Carlo evidence illustrates the outlier problem. Consider a DGP given by:

$$x_{i,t} = \nu_{i,t} \qquad \nu_{i,t} \sim \mathsf{IN}[0,1] \qquad \text{for } i = 1, \ldots, 4. \tag{3.138}$$

We shall generate nonlinear functions given by the inverses of these normal distributions:

$$x^{-1}_{i,t} = \frac{1}{x_{i,t}}. \tag{3.139}$$

The GUM, which contains twenty irrelevant variables, is given by:

$$x^{-1}_{i,t} = \alpha_0 + \sum_{k=1}^{4} \alpha_{i,t-k} x^{-1}_{i,t-k} + \sum_{j=1}^{4} \sum_{(j \neq i)} \sum_{m=0}^{4} \alpha_{j,t-m} x^{-1}_{j,t-m} + \epsilon_t, \tag{3.140}$$

for $t = 1, ..., T$, with $M = 10,000$ replications.

Equation (3.140) leads to $|\mathsf{t}|$-values as large as 19 for variables with zero population non-centralities. The variable would unequivocally, but incorrectly, be retained as a DGP variable. On average, two of the twenty irrelevant regressors are retained at the 1%

significance level. This implies that a fat-tailed distribution would have a null rejection frequency of 10% at the 1% significance level. If the dependent variable is $x_{i,t}$ rather than $x_{i,t}^{-1}$, the retention probabilities are correct. Non-normal errors can also pose a similar problem. Hence, the problem of model selection when there are extreme observations is exacerbated by the inclusion of nonlinear functions such as inverses.

Solution: Indicator saturation techniques

To overcome the problem of fat tails, we draw on the technique proposed by Hendry *et al.* (2004), outlined in Chapter 2, in which the data are saturated with as many indicators as observations and the indicators are selected at a chosen significance level to identify outliers. Once we have identified the outliers we can remove them, and the selection process will not be biased in favour of nonlinear functions that are proxying indicators for the outliers. Crucially, the indicators will not be retained if nonlinearity is the correct LDGP.

This technique also overcomes the problem of undetectable outliers. One concern with nonlinearity is that it is difficult to distinguish between extreme observations that are outliers and extreme observations that are due to the nonlinearity in the data. Methods that remove extreme observations could be in danger of removing the underlying nonlinearity that should be modelled. Indicator saturation techniques can avoid this problem by including all potentially relevant variables as well as indicators for all observations in the initial GUM. By removing the extreme observations in conjunction with selecting the nonlinear functions we avoid the problem of removing observations that generate the nonlinearity.

3.4.5 Super-conservative strategy

Irrelevant nonlinear functions are likely to be detrimental to both modelling and forecasting, hence, nonlinear functions should only be retained if there is definite evidence of nonlinearity because these models are much less robust than linear models, both to changes in collinearity between regressors and location shifts within the equation or in any retained but irrelevant variable. Given a preference for linear models unless strong

evidence for nonlinearity is presented, and the possible excess retention of irrelevant functions due to the large number of nonlinear functions tested, we propose a 'super-conservative' strategy for PcGets. This strategy would use more stringent critical values for the nonlinear functions compared to the linear functions, which would all be tested within the same procedure. Hence, diagnostic tests would apply to the full GUM, but pre-search tests and multipath search tests would be conducted at more stringent critical values for the nonlinear functions. The critical value would depend on the number of functions included in the model and, as with all significance levels, the choice will depend on the preferences of the econometrician.

Block F-tests on classes of nonlinear functions could be incorporated into the pre-search stage in PcGets. Tight significance levels would again be used, and a sequential testing procedure on classes of nonlinear functions entering the GUM would be undertaken until just those classes that are significant are retained to formulate the GUM. This would narrow down the number of nonlinear functions in the multipath search stage.

3.5 Conclusion

This chapter develops a coherent general-to-specific strategy for the selection of nonlinear models which is designed to be embedded within the automatic model selection algorithm of PcGets. First, a GUM is formulated in which all potential variables that are thought to explain the phenomenon of interest are included, assuming a linear functional form. Second, a test of linearity is applied to the GUM to test the linear approximation. The proposed index test is designed to handle large dimensional systems and is therefore appropriate for testing the GUM. The test generates quadratic and cubic functions of all regressors and tests for their joint significance at an appropriate significance level; we propose 1%. If the null is accepted, standard PcGets selection procedures are applied to the linear GUM. However, if the null is rejected, a nonlinear functional form is generated using polynomial transformations of the regressors, and selection follows an alternative procedure.

The PcGets selection procedure for nonlinear models would be implemented as follows. A set of polynomial transformations of the regressors is generated. This would

consist of the quadratic and cubic functions, as in the index test, but could include more general transformations such as cross-products. All polynomial functions are double demeaned prior to inclusion in the GUM to remove potential collinearity. Observe that the index test performs well if there is collinearity, and so orthogonalising transformations are only undertaken at this later stage. Under collinearity, the index test has power against cross-product terms, but as the nonlinear functions are now orthogonalised, inclusion of cross-product terms in the GUM may be preferable. A set of T indicators are also generated and included in the GUM in order to detect outliers concurrently with selection of the specific model, and selection is performed using the techniques developed to handle more variables than observations.

An F-test on all nonlinear functions in the GUM is redundant as the index test has found evidence of nonlinearity. However, F-tests are undertaken on subsets of the nonlinear functions, commencing from the highest order terms, to remove any highly irrelevant nonlinear functions. A potential problem arises if the F-tests on all subsets of nonlinear functions are accepted, contradicting the results of the index test. Hence, we propose the use of a multi-stage strategy, analogous to that implemented in the PcGets pre-search stage, in which tests are undertaken at consecutively tighter significance levels. We recommend commencing with the significance level used in the index test, and so if a 1% level was used in the test, we would propose using a 0.01 significance level, followed by a 0.005 significance level, and finally a 0.001 significance level. It is feasible that the p-value ellipsoid for the 0.01 level does not nest the origin, whereas the ellipsoid for the 0.001 level does, and the resulting model after the pre-search stage must retain nonlinear functions to correspond to the findings of the index test. The multipath search is then implemented. If the general model contains a large number of nonlinear functions, i.e., the pre-search stage did not substantially reduce GUM, the super-conservative strategy is proposed for the nonlinear functions. Again, the final model must contain nonlinearities to justify the results of the index test, and so the multi-stage strategy is implemented within the multipath search procedure. Rules for the super-conservative strategy would be similar to those implemented for the Schwarz information criteria (see Campos *et al.*,

2003), and the selection strategy should deliver the undominated, congruent, specific, nonlinear model. A comprehensive Monte Carlo simulation study across a wide range of states of nature is required to calibrate the super-conservative strategy.

Solutions to three potential difficulties are proposed. Collinearity is one of the most fundamental problems with model selection. We show the correlation magnitudes between linear and nonlinear functions can be extremely high, causing selection algorithms to struggle to identify the relevant variables. This is due to non-zero means in the data. We propose a solution of de-meaning both the linear term prior to the transformation, and after the nonlinear function has been generated. This removes an important component of the collinearity, and PcGets is shown to have good selection properties with this partial orthogonalisation. We emphasise the importance of normality for model selection, and show that the null rejection frequency can be greatly increased relative to the nominal significance level if nonlinear functions capture extreme observations: the solution of indicator saturation is proposed. Further, both this, and the many nonlinear functions created, together always necessitate the use of the multi-stage selection process described earlier for handling more variables than observations. Finally, the potential problem of excess retention of irrelevant variables is addressed, and a super-conservative strategy is proposed. The chapter has shown that it is essential to implement all four of these developments to achieve a successful algorithm, and that removing any one component can be seriously deleterious. Since previous modelling of nonlinearities has not done so, many empirical studies merit a revisit to check their validity.

PART II

NON-STATIONARITY

Chapter 4

Modelling Quarterly UK Inflation, 1967q1-1998q2

4.1 Introduction

The importance of a clear understanding of the determinants of inflation is undisputed given its role in economic policy, and yet there are a plethora of theories regarding its determination. Hence, there is an obvious role for model selection techniques in ascertaining the most relevant causes of inflation. Given the range of plausible explanations, a *Gets* approach that tests the relevance of all possible causes would be the most appropriate methodology to adopt, and this approach is embodied in the automatic model selection algorithm of PcGets, assessed in Chapters 2 and 3. This chapter develops a dominant congruent model of inflation within a dynamic, single-equation framework that encompasses almost all relevant theories and Chapter 5 assesses the model's forecasting performance. The application provides the basis for an assessment of *Gets* model selection techniques, both for in-sample and forecasting models, in an empirical setting. Furthermore, as UK inflation has experienced many structural breaks and regime changes over the past 40 years, we test the modelling and forecasting techniques used against such non-stationarities.

The key determinants of inflation are the pressures arising from excess demands in all sectors of the economy and so measures of excess demand for both goods and labour are developed.[1] The output gap, as a proxy for excess demand for goods and services, is widely used but is notoriously difficult to measure due to it being a latent variable. Three measures of the output gap are derived to establish robustness of the measure used. First, a standard production function (PF) approach is undertaken. Second, an alternative PF approach is proposed, modelling the production function within a dynamic framework in

[1] Hendry (2001) develops excess demands for both money and debt but finds them to be insignificant in explaining inflation, and so we exclude these measures from our analysis.

which total factor productivity (TFP) is modelled as a random walk with drift. Third, a split trend is used to capture TFP in a simple Solow residual framework. In theory the gaps should be identical, but there are some differences due to measurement error, and the Solow residual gap is used in the resulting analysis. The excess demand for labour measure is based on the disequilibrium unemployment generated when the real interest rate exceeds the real growth rate and vice versa.

A congruent model of quarterly UK inflation is developed in which most extant theories of inflation play a role in its determination, including unit labour costs, import prices, producer prices, exchange rates, foreign prices, and excess demand for both goods and labour. The model selection technique allows all individual variables to partially explain inflation. This is based on the assumption that the DGP is a complex distribution function, $\mathsf{D}_{\mathbf{u}}\left(\mathbf{U}_T^1|\mathbf{U}_0,\boldsymbol{\psi}_T^1\right)$, of all potential variables in the economic mechanism, but it can be reduced to a LDGP via the theory of reduction, see Chapter 1. The LDGP consists of the relevant variables for the phenemenon under analysis, see Hendry (1995, ch.9), which is mapped from a joint distribution function to an empirical model. However, under an alternative paradigm, the marginalization and conditioning required in the theory of reduction are not imposed. Instead of individual variables contributing to inflation, there are underlying generic factors capturing the information in $\mathsf{D}_{\mathbf{u}}\left(\mathbf{U}_T^1|\mathbf{U}_0,\boldsymbol{\psi}_T^1\right)$. Stock and Watson (1998, 1999a, 1999c) propose that inflation is driven by general business cycle factors that capture all demand and supply pressures in the economy. We test whether there is a generic 'business cycle' factor driving inflation, but find that evidence is limited. A business cycle component does not negate the individual variables included, and very few principal components are retained, with little explanatory power. The lack of interpretability and non-robustness to changes in the information set further limit the use of a broad cyclical factor.

The chapter is structured as follows. Section 4.2 discusses the data, developing measures of the output gap in detail. Section 4.3 presents the model of quarterly UK inflation. Section 4.4 assesses whether there is a generic business cycle component to inflation, and finally, section 4.5 concludes.

4.2 Data

The data set consists of quarterly data for the UK over 1965q1-2003q2 and is derived from a number of sources detailed in Appendix 4.A. All data are seasonally adjusted and lower cases represent logarithms. Indicator variables, given by I_{date}, take the value 1 in the quarter indexed and 0 otherwise. Dummy variables combine two indicators, taking the values -1 and +1 in the indexed quarters. The estimation sample period for the inflation model is 1967q1-1998q2, resulting in 126 observations, with 20 observations retained for the forecasting exercise undertaken in Chapter 5, extending from 1998q3-2003q2. Restricting the sample endpoint in this chapter to 1998q2 enables *ex ante* forecasts to be computed, thereby avoiding any bias in the forecasts due to a preliminary examination of the out-of-sample data. The excess demand measures are estimated over 1966q2-2003q2 to facilitate forecasts over the full forecast horizon.

The order of integration of price level data has been discussed extensively in the literature. Hendry (2001) concludes that the price level is I(1) but contains deterministic shifts, giving the impression that the series is I(2). Dickey-Fuller tests are rarely conclusive due to their low power, and results differ across countries and time periods. However, the ADF test statistics for the implicit GDP deflator suggest that the price level is I(2) and the inflation rate is I(1).[2] This implies that there are two forms of cointegration. First, the price measures cointegrate to I(1) and secondly, the I(1) cointegrating price measures drive fluctuations in the inflation rate, yielding a polynomially cointegrating relation. This will give a long-run solution for the price level and a long-run solution for the inflation rate based on relative prices. The evidence for deterministic shifts in an I(0) process is also convincing. Labhard, Kapetanios and Price (2005) apply the Bai and Perron (1998) break test to UK CPI inflation over 1984q1-2005q1 and they identify three breaks. Removing these mean shifts results in an I(0) inflation process. If we apply both the Bai and Perron test and the Hendry *et al.* (2004) indicator saturation test to the GDP deflator over the sample 1967q1-2003q2, we find breaks at approximately 1972q4, 1979q1 and 1992q1. The deterministic shifts in the GDP deflator could conceivably co-

[2] ADF test results with constant and trend: $\mathsf{H}_0 = I(1) : ADF\ \tau = -0.906$, $\mathsf{H}_0 = I(2) : ADF\ \tau = -2.884$, $\mathsf{H}_0 = I(3) : ADF\ \tau = -13.82^{**}$.

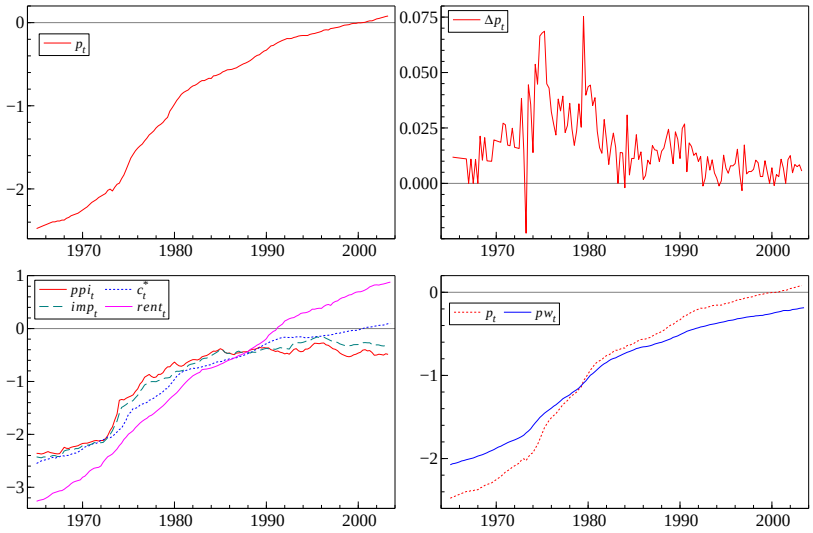

Figure 4.1: GDP deflator; quarterly inflation rate; producer prices, unit labour costs, imports and housing rent; and the external and internal price levels

break with the explanatory variables, resulting in an I(0) model. Note that many studies examine the consumer price deflator or the net national income deflator as opposed to the GDP deflator. Hendry (2001) finds that these series do not mutually cointegrate, and so empirical models are specific to the price measure used. Figure 4.1a records the log of the GDP deflator (p) and panel b records the quarterly inflation rate (Δp). There is a substantial negative inflation outlier of -2.2% in 1973q2. This is thought to be measurement error as the change in inflation exhibits a blip, with an initial fall and subsequent rise of similar magnitude, hence an indicator variable is included for this outlier.

The equilibrium correction model of inflation is based on a mark-up model, with excess demand pressures causing short-run cyclical movements in inflation while the long-run price level is determined by sectoral price levels. The main series include producer prices (ppi), import prices (imp), housing rent ($rent$), unit labour costs scaled for the decline in average hours (c^*), oil prices (oil), national debt (n) and external prices (pw). The short-run pressures are captured by the output gap $\left(y^d\right)$, excess demand for unemployment $\left(U^d\right)$, the growth rate of broad money ($\Delta m4$), the short-term interest rate (R_s), the long-term interest rate (R_l), the short-long real interest rate spread (s), and the real effective exchange rate (e_r).

4.2.1 Sectoral prices

Figure 4.1c records the levels of ppi, imp, c^*, and $rent$, and Figure 4.2 records the quarterly growth rates of ppi, imp, c^*, and $rent$ in panels a to d respectively, recorded with the inflation rate for comparison. Δppi follows price inflation fairly closely, but both Δimp and $\Delta rent$ are much more volatile than inflation. Housing market volatility has increased substantially since the late 1980s boom and subsequent recession. Unit labour costs for the whole economy are adjusted for the gradual decline in average working hours by 0.0625% per quarter from 1965q1-1984q4, and for a decline of 0.03% hours per quarter from 1985q1-2003q2. This accords with the data on normal hours taken from the national statistics figures on average actual weekly hours of work. Data limitations prevent a more disaggregated approach that also controls for the effects of self-employment and for the slower evolvement of wage-price linkages in the public sector.

Figure 4.3a records the real oil price scaled for a zero mean $(oil - p)$. The 1973 and 1979 oil price shocks are evident, as are large swings in the late 1990s. Real unit labour costs $(c^* - p)$ are recorded in panel b, and real producer prices are recorded in panel c, exhibiting a systematic decline since the mid-1970s. Finally, panel d records the mark-up, π^*, derived in (4.19) as the profit over and above unit labour costs, producer prices, and world prices in sterling.

4.2.2 World prices and exchange rates

Theories of inflation based on purchasing power parity (PPP) argue that in the long-run, exchange rates should adjust to eliminate arbitrage opportunities, and hence inflation will be imported via pass-through effects. The external price level is based on the OECD consumer price index and therefore will not be fully external as the measure will contain UK inflation. Furthermore, the measure will exclude important growth countries such as China and India. $pw_{\pounds,t} = pw_t - e_t$, where e_t is the nominal exchange rate, denotes world prices in sterling, and therefore e_r is derived (setting the sample mean to zero) as:

$$e_{r,t} = p_t - pw_{\pounds,t} + 0.02. \tag{4.1}$$

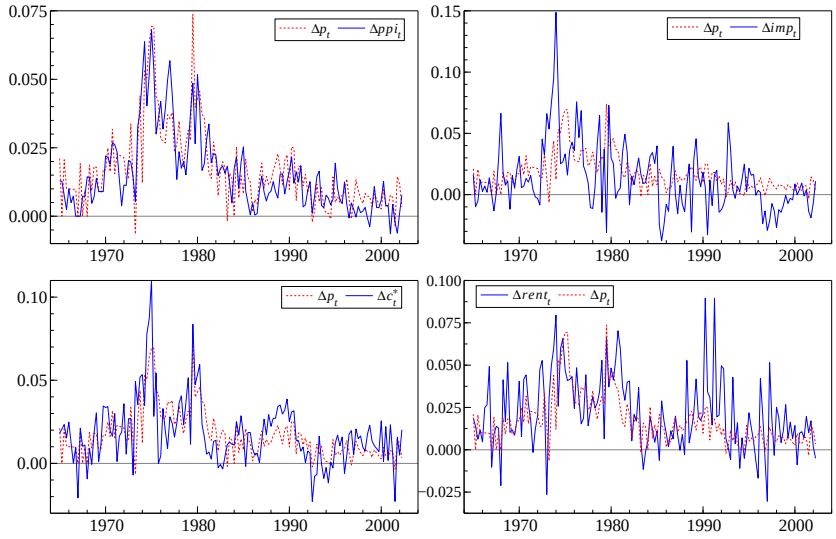

Figure 4.2: Quarterly growth rates of the producer price index; import prices; unit labour costs; and housing rent

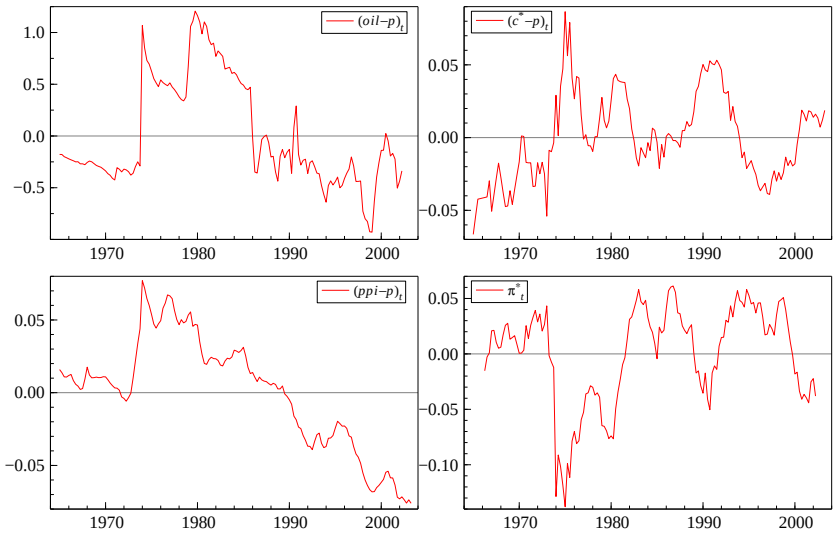

Figure 4.3: Real oil prices; real unit labour costs; real producer prices; and the markup

Figure 4.1d records the external price level against the UK price level for comparison: UK prices have grown faster on average. Figure 4.5a records the external inflation rate which shows that patterns of inflation are similar to those for the UK.

Figure 4.4, panels a and b, record the nominal and real exchange rates respectively. There are substantial and persistent deviations from PPP over the period, with a range extending from +20% to -30%. While e_r is judged to be I(0) over very long data sets, the ADF statistics for the period 1965q1-2003q2 find e_r to be I(1).

4.2.3 Interest rates and monetary aggregates

The short-long real interest rate spread captures the inflationary pressures arising from an increase in the cost of capital relative to the borrowing rate, assuming the short rate is the control variable and the long rate is a proxy for the cost of capital. As the interest rates are annual measures, the spread is scaled to represent a quarterly measure and is adjusted for a zero sample mean. The spread is recorded in Figure 4.4c.

Monetary theories of inflation stem from the 'quantity theory' of Friedman (1956), in which money is treated as exogenous, enabling the money demand equation to be inverted in order to solve for the price level. There is a vast literature looking at money causing inflation, but Hendry (2000a) finds no support for this theory. The growth rate of broad money, recorded in 4.4d, tends to exceed price inflation over the 1980s as people transferred their holdings from narrow money to broad money due to the tightening operated via the Medium Term Financial Strategy, and again over the latter part of the 1990s. A measure of excess money holdings is not developed but the inclusion of the growth rate of broad money does test this theory in the inflation model.

4.2.4 Excess demand for labour

There is a substantial literature examining the importance of labour market pressures on inflation. We use a measure of excess demand for unemployment based on an equilibrium correction model, which models the change in the unemployment rate, $\Delta U_{r,t}$, as a function of the gap between the real interest rate and the real growth rate, following Hendry (2001). In this model, disequilibrium unemployment is based on steady state growth,

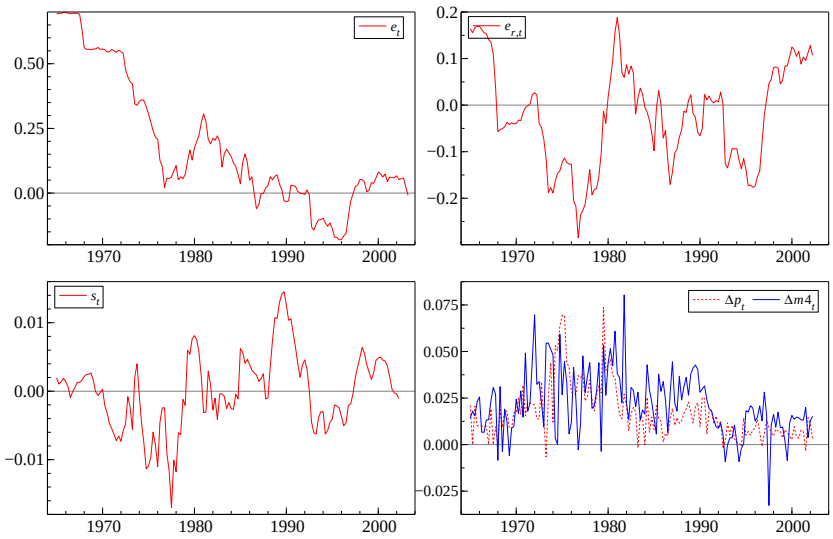

Figure 4.4: The nominal effective exchange rate; real effective exchange rate; short-long interest rate spread; and the growth rate of broad money

with unemployment rising when the real interest rate exceeds the real growth rate and vice versa. Disequilibrium unemployment is described as:

$$\begin{aligned}
\Delta U_{r,t} &= \underset{(0.004)}{0.010}\Delta_4 \left(R_l - \Delta p - \Delta y\right)_t + \underset{(0.034)}{0.869}\Delta U_{r,t-1} \\
&\quad + \underset{(0.003)}{0.008}\left(R_l - \Delta p - \Delta y\right)_{t-1} - \underset{(0.003)}{0.010}U_{r,t-1} - \underset{(0.001)}{0.005}I_{71:1} \\
&\quad + \underset{(0.001)}{0.007}I_{71:2} - \underset{(0.001)}{0.003}I_{76:1} + \underset{(0.001)}{0.003}I_{90:3} + \underset{(0.001)}{0.003}I_{91:1} \qquad (4.2) \\
\mathsf{R}^2 &= 0.860 \quad \widehat{\sigma} = 0.110\% \quad SIC = -13.524 \quad \mathsf{F}_{\mathsf{ar}}(5,135) = 0.924 \\
\mathsf{F}_{\mathsf{arch}}(4,132) &= 2.684^* \quad \mathsf{F}_{\mathsf{het}}(13,126) = 1.379 \quad \chi^2_{\mathsf{nd}}(2) = 0.475 \\
\mathsf{F}_{\mathsf{reset}}(1,139) &= 0.204 \quad \mathsf{F}_{\mathsf{Chow}}(10,130) = 0.445 \quad T = 1966q2 - 2003q2.
\end{aligned}$$

The model provides a reasonable fit and passes all diagnostics apart from ARCH at 5% significance. Five indicators are included in the model but do not enter into the long-run solution. The resulting excess demand for labour measure is given by:

$$U^d_t = U_{r,t} - 0.73\left(R_{l,t} - \Delta p_t - \Delta y_t\right). \qquad (4.3)$$

Figure 4.5 records the quarterly change in the unemployment rate and excess demand for labour in panels b and c respectively. The excess demand for labour measure follows a similar pattern to measures of the UK NAIRU estimated by, *inter alia*, Batini and

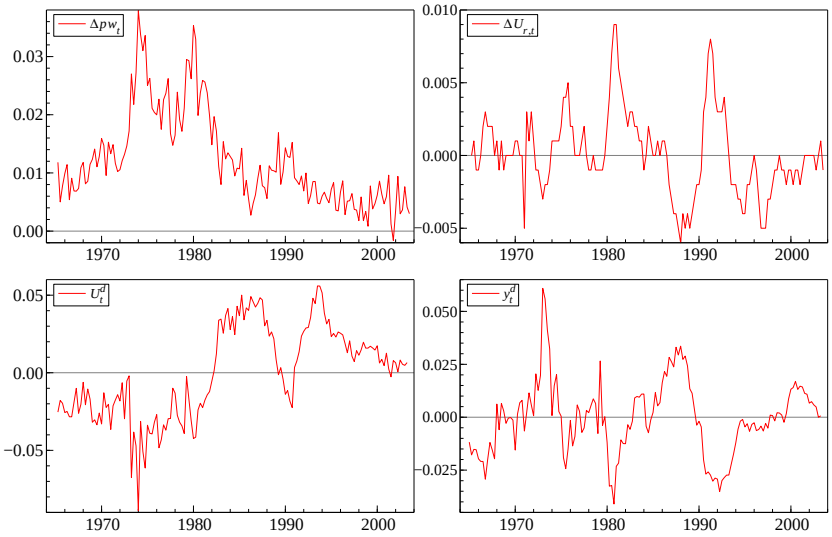

Figure 4.5: The external quarterly inflation rate; the quarterly change in the unemployment rate; excess demand for labour; and excess demand for goods and services

Greenslade (2006).

4.2.5 Excess demand for goods and services

One of the fundamental driving forces of inflation is excess demand for goods and services, but it is extremely difficult to measure accurately, see De-Brouwer (1998) and Hendry (2000c). This section derives estimates of the gap based on a static production function, a dynamic production function and a Solow residual measure with split time trend to capture productivity gains. The third measure is used in the subsequent inflation model.

Static production function

Assuming a Cobb-Douglas technology with constant returns to scale, an elasticity of substitution equal to unity, and Hicks-neutral productivity, the production function is given as:

$$Y_t = A_t \overline{L}_t^{\alpha} K_t^{1-\alpha}, \tag{4.4}$$

where $\overline{L}_t$ is labour input, K_t is capital input, A_t is total factor productivity (TFP), or the efficiency with which both capital and labour are used to produce output, and α is the elasticity of output with respect to labour ($0 < \alpha < 1$). Under the assumption of perfect competition, where the marginal products of labour and capital are equal to the wage

rate and profit rate, α is equal to the share of output going to labour, or equivalently the share of wages (W_t) in total income (Z_t). Over the sample period α lies in the range $[0.68, 0.72]$, with a mean of 0.702. Hence, we set $\alpha = 0.7$.

$\overline{L}_t$ comprises employment (L_t) and the number of paid hours worked per employee (H_t). L_t is broken down into three components:

$$l_t = wpop_t + part_{r,t} + emp_{r,t}, \tag{4.5}$$

where $wpop_t$ is the population of working age, $part_{r,t}$ is the participation rate and $emp_{r,t}$ is the employment rate. $\overline{L}_t$ should be adjusted for labour quality but such data is unavailable and so any changes to labour quality will be picked up in the residual. l_t is recorded in Figure 4.6a, along with trend employment. H_t is approximated by the difference between average overtime hours and average undertime hours, but as the impact of short-time is negligible, hours can be calculated as:

$$h_t \approx \ln\left[\overline{H}_t\left(1 + OH_t\right)\right], \tag{4.6}$$

where $\overline{H}_t$ is the normal number of hours worked per week and OH_t is the number of overtime hours worked per week. $\overline{H}_t$ declined from 39 hours in 1965 to 32 hours in 2002. The implied fall in output is offset by an increase in efficiency that will be captured in A_t. $\overline{h}_t$ and OH_t are recorded in Figure 4.6, panels c and d respectively, and it is evident that an increase in systematic overtime has accompanied the fall in normal hours. Muellbauer (1984) finds that data on average hours provides a good approximation to labour utilisation.

Capital input (K_t) is measured by the net capital stock excluding the dwellings sector (J_t). This is a wealth measure of capital that weights different types of capital by their asset prices. The theoretical concept requires a measure of capital services that captures the flow of productive input from capital. Thus, we assume that capital services are proportional to the asset value measure of capital stock. If capital is thought of as an overhead, J_t may not need to be adjusted for the degree of capacity utilisation, $U_{c,t}$. However, $U_{c,t}$ substantially reduces the procyclicality of the residual, and so capital input

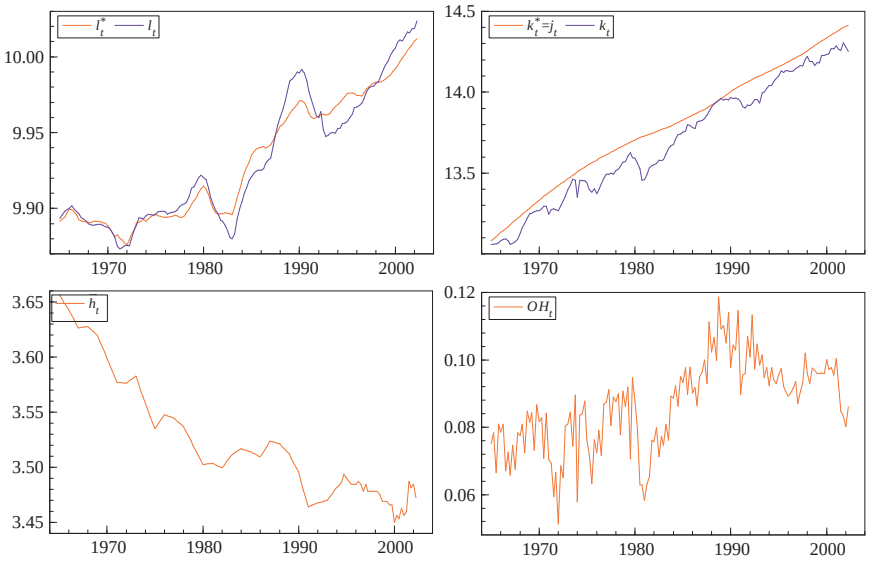

Figure 4.6: Employment; capital stock and utilisation of capital; normal number of hours worked per week; and number of overtime hours

is measured as:

$$k_t = j_t + U_{c,t}. \tag{4.7}$$

k_t and j_t are recorded in Figure 4.6b. The capacity utilisation measure is constructed from the CBI industrial trends survey in which firms report whether they are operating below normal capacity levels, see the Appendix for details. Full capacity is assumed to use approximately 91% of the total capital stock available. The data applies to manufacturing output, but as services have increased dramatically over the period of estimation and the relationship between utilisation rates for manufacturing and services is ambiguous, the utilisation measure may be a poor approximation. A shortage of data on capacity utilisation levels in the service sector prevents a more rigorous, disaggregated measure being derived.

Estimates of the latent variables including potential capital, potential labour, and TFP (denoted by superscript *) are required to calculate potential output. For capital input, it is assumed that capital operates at full capacity, hence $U_{c,t} = 0$ and $k_t^* = j_t$. Even though net investment per annum is very volatile, it is such a small fraction of net capital stock as to have a very limited impact on the stock of capital.

The working population is assumed to be at trend. Most movements in the working population are long-run or permanent changes caused by, for example, a change in pen-

sion provisions, changes in the age of retirement, or an increase in the number of women who work. There may be a small cyclical component to the working population, for example, in the climate of a recession some members may choose to remove themselves from the working population pool by retiring early or choosing not to search for a job, but it is assumed that this effect is negligible. The trend employment rate is derived from the trend unemployment rate, which is used as a proxy for the NAIRU. This is calculated using the unobserved components (UC) method of decomposition based on a stochastic level and cycle.[3] To estimate the trend participation rate, the total number in employment is smoothed using a Hodrick Prescott (HP) filter, and the level of unemployment is derived from the trend unemployment rate.[4] The sum is then divided by the actual working population to result in the potential participation rate. $\overline{H}$ is assumed to pick up long-run trends only, as any cyclical fluctuations will not be captured due to labour hoarding. Also, overtime hours are assumed to be 0, therefore $h_t^* = \overline{h}_t$.

The calculation of trend a_t depends on the assumptions made regarding the nature of TFP growth. The most appropriate method for detrending a_t depends crucially on whether technical innovations are thought to be random shocks due to a burst of new ideas, or whether ideas diffuse gradually as learning is slowly accumulated. One may expect productivity shocks to take their time feeding through as the learning process, along with research and development, occurs. Also, shocks that are specific to sectors are likely to only impact gradually in the aggregate. Hence, a plausible trend would be smooth but would also allow for random productivity shocks. a_t will also pick up

[3]The univariate UC model is given as:

$$\begin{aligned} y_t &= \mu_t + \psi_t + \varepsilon_t \\ \mu_t &= \mu_{t-1} + \beta_{t-1} + \eta_t \\ \beta_t &= \beta_{t-1} + \zeta_t \\ \begin{bmatrix} \psi_t \\ \psi_t^* \end{bmatrix} &= \rho \begin{bmatrix} \cos\lambda_c & \sin\lambda_c \\ -\sin\lambda_c & \cos\lambda_c \end{bmatrix} \begin{bmatrix} \psi_{t-1} \\ \psi_{t-1}^* \end{bmatrix} + \begin{bmatrix} \kappa_t \\ \kappa_t^* \end{bmatrix} \end{aligned}$$

where $t = 1, ..., T$. λ_c is the frequency in radians, $0 < \lambda_c < \pi$, ρ is the dampening factor, $0 < \rho \leq 1$, and $\varepsilon_t \sim \mathsf{IN}\left[0, \sigma_\varepsilon^2\right]$, $\eta_t \sim \mathsf{IN}\left[0, \sigma_\eta^2\right]$, $\zeta_t \sim \mathsf{IN}[0, \sigma_\zeta^2]$ and $\kappa_t, \kappa_t^* \sim \mathsf{IN}\left[0, \sigma_\kappa^2\right]$. The disturbances of each of the components are assumed to be mutually uncorrelated. See Koopman *et al.* (1995) for a detailed outline.

[4]The HP filter is a two-sided symmetric moving average filter that decomposes a series into a trend and cycle by optimising:

$$y_t^* = \arg\min_{y_t^*} \sum_{t=1}^{T} (y_t - y_t^*)^2 + \lambda \sum_{t=3}^{T} \left(\Delta^2 y_t^*\right)^2,$$

where λ is the smoothness parameter. For quarterly data, λ=1600 is used as a rule of thumb.

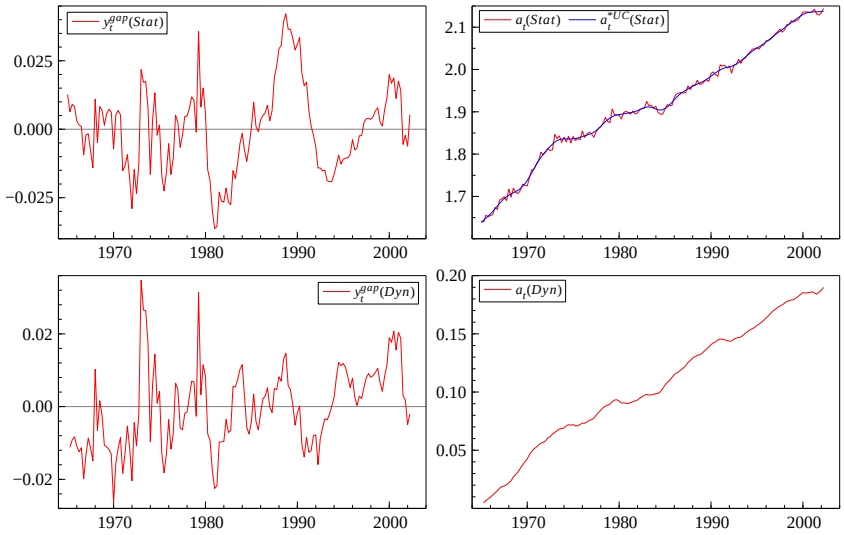

Figure 4.7: The static production function output gap; TFP estimated by the static production function; the dynamic production function output gap; and TFP estimated by the dynamic production function

efficiency gains in the quality of capital and labour. Given these assumptions, a UC model is used to detrend a_t based on a smooth trend.

The resulting output gap, defined as $y_t - y_t^*$, where $y_t^* = a_t^* + \alpha \bar{l}_t^* + (1-\alpha)\, k_t^*$, is recorded in Figure 4.7a. The 1980s recession is estimated to be a lot deeper than the 1990s recession, reaching a magnitude of 3.6% compared to 2.0% of output in the early 1990s. One explanation for the smaller recession in the 1990s is the sharp drop in normal hours at the beginning of the 1990s that is unlikely to be offset by increasing productivity due to efficiency gains, causing lower potential output and reducing the size of the negative gap. Panel b records a_t and the smoothed estimate based on a UC model with fixed level and stochastic slope, a_t^{*UC}. The actual and trend estimates are very similar, and the lack of cyclicality in TFP shows that the utilisation rates have accounted for business cycle fluctuations. Productivity declined following the first oil price shock but then picked up again in the mid-1980s. Productivity does flatten off from 2000 onwards, although caveats regarding end-of-sample estimates from such decomposition techniques apply.

Dynamic production function

The production function is a static and cointegrating concept. Hence, the standard growth accounting framework should be sufficient. However, the presence of substantial

measurement errors in K, $\overline{L}$, and A imply that a stable relationship may be difficult to identify. Haavelmo (1944) highlights the problem by distinguishing between the latent variables identified in economic theory, their correctly measured empirical counterparts, and the actual data available that contains substantial measurement error. With regard to capital, theoretical models require a measure of the flow of capital services in the economy, but empirical data is a measure of the capital stock, containing errors due to the assumptions made about depreciation, scrapping, aggregation, etc. In order to allow for measurement error, we shall analyse the PF in a log-linear dynamic setting, resulting in a stable solution for potential output. This approach has the added advantage of setting the PF in the long-run context. Firms do not produce to the PF constraint on a short-run basis. The magnitude and volatility of inventories highlight this fact. In the short-run, firms tend to produce to inventory or order and then sell from these. However, in the long-run the PF constraints will bite, so a dynamic model that allows for adjustments over the short and medium term is appropriate.

The dynamic PF model is set in the single-equation framework with a time-varying regression intercept that captures unobserved TFP, and is augmented by I(0) cyclical factors. The long-run solution, proxying potential output, will be based on the static PF model:

$$y_t^* = \Psi_t + \gamma_1 k_t + \gamma_2 \overline{l}_t, \tag{4.8}$$

where Ψ_t is a local level with drift intercept term capturing a_t. We assume that a single-equation analysis of Δy_t is valid, requiring $\overline{l}_t$, and k_t to be weakly exogenous for y_t. Given an ADL(1,1) model:

$$y_t = \psi a_t + \beta_1 y_{t-1} + \beta_2 k_t + \beta_3 k_{t-1} + \beta_4 \overline{l}_t + \beta_5 \overline{l}_{t-1} + \delta' \left(\text{cyclical factors}\right) + \varepsilon_t, \tag{4.9}$$

where $\varepsilon_t \sim \mathsf{IN}\left[0, \sigma_\varepsilon^2\right]$, we can estimate the model in EqCM form:

$$\begin{aligned} \Delta y_t &= \psi a_t + \beta_2 \Delta k_t + \beta_4 \Delta \overline{l}_t + (\beta_1 - 1)\left[y_{t-1} - \kappa_1 k_{t-1} - \kappa_2 \overline{l}_{t-1}\right] \\ &\quad + \delta' \Delta \left(\text{cyclical factors}\right) + \varepsilon_t, \end{aligned} \tag{4.10}$$

where $\kappa_1 = \frac{\beta_2 + \beta_3}{1 - \beta_1}$ and $\kappa_2 = \frac{\beta_4 + \beta_5}{1 - \beta_1}$. The time-varying intercept evolves according to the

transition equation:

$$a_t = a_{t-1} + \mu + \eta_t, \qquad \eta_t \sim \mathsf{IN}\left[0, \sigma_\eta^2\right]. \tag{4.11}$$

We assume σ_ε^2 and σ_η^2 are independently distributed. The model is formulated in state space form and is estimated using the Kalman (1960) Filter. The time-varying trend, modelled as a random walk with drift, allows for permanent shifts in TFP. This will robustify the coefficient estimates against the effects of structural change. The time-varying trend will proxy advances in human capital, including knowledge accumulation, experience and educational improvements. Human capital is captured by the process of cohort arrival and departure in the labour force. Retirements from the workforce tend to have a lot of experience but education occurred a long time ago, whereas new arrivals have a recent education but a lack of experience. Aggregating across all individuals, given that workers are at different stages in their lifecycles, implies a smooth growth in the effective labour force. Furthermore, a random walk with drift will capture the effect of human capital using a physical capital stock that embodies technological progress.

Equation (4.10) is generalised to allow for a broader dynamic structure, which is identified using a *Gets* modelling strategy. The GUM consists of the lagged levels of output, capital and labour, contemporaneous values and four lags of Δk and $\Delta \overline{l}$, cyclical factors including the change in overtime hours (ΔOH), change in capacity utilisation (ΔU_c), change in inventories, ($\Delta invent$), real interest rates (R_l^r and R_s^r) the real exchange rate (e_r), and a convex investment adjustment cost given by $\left(I_t^2/K_t\right)$, and impulse and blip dummies for the large outliers. The model is estimated using STAMP (Koopman *et al.*, 1995) and the *Gets* reduction is implemented manually to mimic PcGets. The resulting model is given by:

$$\begin{aligned}
\Delta y_t &= \underset{(1.90)}{1.103} a_t + \underset{(0.0013)}{0.002}\,\mu - \underset{(0.06)}{0.413} y_{t-1} + \underset{(0.05)}{0.098} k_{t-1} + \underset{(0.11)}{0.302} \overline{l}_{t-1} \\
&\quad + \underset{(0.25)}{0.555} \Delta \overline{l}_t + \underset{(0.007)}{0.031} \Delta invent_t + \underset{(0.05)}{0.156} \Delta U_{c,t} \\
&\quad + \underset{(0.006)}{0.027} I_{68:1} + \underset{(0.006)}{0.035} I_{73:1} + \underset{(0.01)}{0.027} D_{79:2} \qquad (4.12) \\
\mathsf{R}^2 &= 0.776 \quad \widehat{\sigma} = 0.664\% \quad \chi^2_{\mathsf{nd}}(2) = 1.347 \quad \mathsf{Q}_{\mathsf{bl}}(11,10) = 12.088 \\
\mathsf{SC}_{\mathsf{r}}(1) &= 0.020 \quad \mathsf{F}^{\mathsf{np}}_{\mathsf{het}}(49,49) = 0.095 \quad \mathsf{DW} = 1.986 \quad T = 1966q2 - 2003q2.
\end{aligned}$$

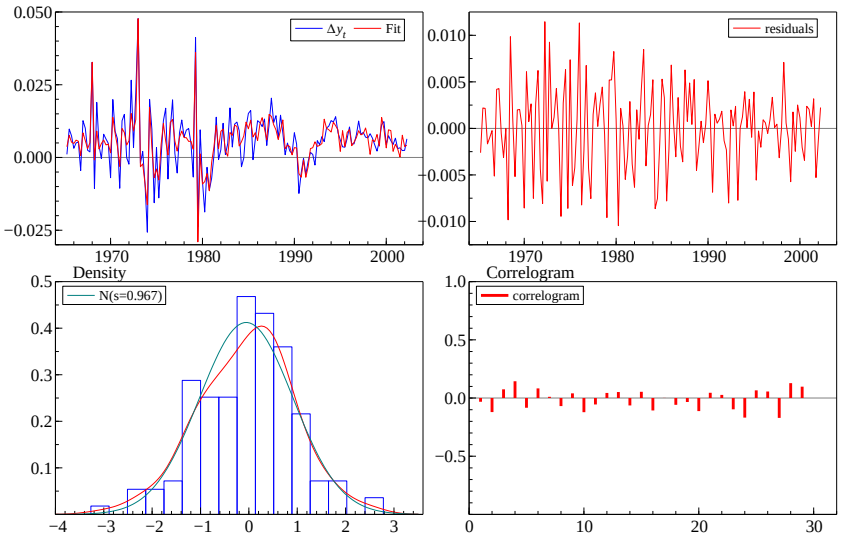

Figure 4.8: Actual and fitted values, residuals, density and correlogram for the dynamic production function

The model represents a good fit with an equation standard error of 0.66% and the model passes all diagnostics. Figure 4.8 records the actual and fitted values, residuals, density, and correlogram. The resulting model is remarkably parsimonious given the number of explanatory variables in the GUM. The random walk and drift components are insignificant, with a very large standard error on a_t, due to collinearity. Hence, we seek a more orthogonal representation in equation (4.13).

The model has an adjustment coefficient of 0.41, implying that two fifths of the disequilibrium at $t-1$ is removed in the following quarter. The current dated adjustment in k is insignificant and so all adjustment to capital takes place in the error correction term. The adjustment term on $\Delta\overline{l}_t$ is large. In period t, firms will not only consider whether they were in equilibrium last period but also whether there is a change in labour input in the current period, and so current decisions have a direct impact on Δy_t. $D_{79:2}$ is a blip dummy taking the values 1 in 1979q2 and -1 in 1979q3, and hence integrates to an impulse dummy which does not enter into the long-run solution. $I_{68:1}$ and $I_{73:1}$ are impulse dummies but are not persistent. Instead, they are capturing one-off shocks or outliers and should not enter the long-run solution as level shifts.[5] The time-varying

[5]The indicators are not included as blip dummies because the counteracting residuals do not occur in the immediate quarter following the positive shock but over the following year: summing the residuals over the following four quarters removes the majority of the shock.

trend will capture the persistent shocks to output.

Equation (4.12) suggests that we can impose a restriction of constant returns to scale. Reparameterising the model to derive output per unit of labour results in a cointegrating relation with coefficients of 0.7 on labour and 0.3 on capital. The model is given by:

$$\begin{aligned}
\Delta\left(y-\overline{l}\right)_t &= \underset{(0.557)}{0.975}a_t + \underset{(0.0004)}{0.001}\,\mu - \underset{(0.055)}{0.375}\left(y-0.7\overline{l}-0.3k\right)_{t-1} + \underset{(0.007)}{0.027}\Delta invent_t \\
&\quad + \underset{(0.053)}{0.137}\Delta U_{c,t} + \underset{(0.007)}{0.025}I_{68:1} + \underset{(0.007)}{0.030}I_{73:1} + \underset{(0.005)}{0.027}D_{79:2} \qquad (4.13)\\
\mathsf{R}^2 &= 0.743 \;\; \widehat{\sigma} = 0.70\% \;\; \chi^2_{\mathsf{nd}}(2) = 4.016 \;\; \mathsf{Q}_{\mathsf{bl}}(11,10) = 5.095. \\
\mathsf{SC}_{\mathsf{r}}(1) &= 0.019 \;\; \mathsf{F}^{\mathsf{nd}}_{\mathsf{het}}(49,49) = 0.163 \;\; \mathsf{DW} = 1.992 \;\; T = 1966q2 - 2003q2.
\end{aligned}$$

The model passes all diagnostics and the equation standard error is only marginally increased to 0.7%. The q-ratio, determined as the ratio of the variance of the unobserved component to the variance of the model residuals, is 0.14. The parameters are relatively stable when imposing the restriction and the drift of 0.1% is now significant. Furthermore, the standard error of a_t has dramatically reduced due to the orthogonal representation, with TFP entering the short-run dynamic model with a near unit coefficient resulting in growth of approximately 20% over the period of estimation. The long-run solution determines the total growth in TFP over the period.

Figure 4.7d records the estimated local level, proxying TFP. The productivity slowdown in the 1970s, the increase in the second half of the 1980s, and the 'new economy' productivity increases of the late 1990s correspond closely to those of the static estimates in panel b. Again, evidence of a slowdown is observed from 2000. The static estimate of TFP is smoother than the random walk with drift, and this is reflected in a trend that tracks output more closely in the dynamic case, resulting in a smaller gap.

The long-run solution is given as:

$$y^*_t = 0.3k_t + 0.7\overline{l}_t + 2.6a_t. \qquad (4.14)$$

The coefficients on capital and labour match those of the static model, in which strong assumptions were made to derive estimates of α. Figure 4.7c records the output gap from the dynamic PF. While the gaps in the 1970s match those of the static PF gap, the late

1980s boom and early 1990s recession are estimated to be much smaller in the dynamic setting. The shocks in the 1970s are quite clearly attributable to short-run shocks and so are not picked up in the long-run trend, whereas the local level component estimates a slowdown in productivity between 1988 and 1992 that is not picked up in the residual based estimation of TFP to the same extent. The static estimates of TFP growth are approximately 50% over the period. A comparable TFP growth of approximately 53% is found in the dynamic model. Output growth between 1966 and 2003 is approximately 130% and so TFP growth accounts for about two fifths of output growth over the period. This is plausible given the large increases in labour participation over the period of estimation.

Solow residual method

The final measure of excess demand overcomes some of the difficulties in estimating a complex PF with stochastic TFP by focusing on the Solow residual (SR), using deviations from a measure of potential capacity to calculate the gap, in which changes in TFP are captured by a segmented linear trend. To determine the impact of technical progress, the coefficient on capital is set at 0.36, which is marginally larger than the coefficient of 0.30 found in the static and dynamic production approaches but it coincides with the annual analogue in Hendry (2001). Robustness of the results is confirmed by varying the coefficient between $[0.3, 0.4]$, which had little impact on the resulting trend estimates. The PF approaches clearly show the trend productivity rate fluctuates over time. A segmented linear trend attempts to capture this, but too many segments will result in a volatile growth rate and sub-sample estimates will be based on very few observations. Separate trends were estimated for the 1960s and 1970s, compared with the 1980s and 1990s. This closely matches the breakpoint found in the early 1980s in the PF approaches, although it does not fully capture the fast growth of the late 1960s, or the acceleration in the late 1990s. The resulting measure of capacity is given in (4.15):

$$cap_t = \begin{cases} 2.53 + 0.0026t + 0.36(k_t - wpop_t) & 1966q2 - 1980q4 \\ 2.46 + 0.0033t + 0.36(k_t - wpop_t) & 1981q1 - 2003q2 \end{cases} \tag{4.15}$$

From this measure of capacity, we calculate excess demand for goods and services as:

$$y_t^d = y_t^l - cap_t, \tag{4.16}$$

where y_t^l is output per worker. The gap measure, recorded in Figure 4.5d, matches the historical record of recessions and booms, both in terms of magnitude and timing, better than the PF approaches. Furthermore, the gap measure is preferable to univariate statistical procedures such as the HP filter. Hence, it is the preferred measure of the output gap used in the inflation analysis below.

Figure 4.9 records all three gap measures for comparison. The volatility of excess demand in the 1970s is captured by all three measures. The movements in the gaps are similar, except for 1971-1973, when the SR gap is positive but the PF gaps are negative. The magnitude of positive excess demand following the 1973 oil shock is much larger for the SR gap, as the stochastic trend measures attribute a proportion of the shock towards higher potential output. The gaps are almost identical over the second half of the 1970s. The early 1980s recession is deepest using the SR and static PF measures, but although they are of similar magnitude, the timing of the static PF trough is approximately one year later. The SR measure shows a faster recovery than the static PF gap, with the late 1980s boom occuring one to two years in advance of the static PF gap. The SR measure has a much deeper recession in the 1990s than either PF measure, reaching a trough of over 3% of GDP. The economy does not recover until the late 1990s, when all three measures estimate a positive gap of approximately 1.5%. The magnitude of the swings for the dynamic PF suggest that changes in TFP are capturing many of the shifts that would be attributed to short-run fluctuations in the two other measures. Despite obvious differences in the estimates of the gaps, there is a consensus as to general excess demand or supply side pressures, and the gap is a key component in explaining inflation.

One could think of the true output gap as a weighted average of the three measures. If each measure of the gap is a combination of the true output gap plus some error:

$$\widehat{y}_{i,t}^{gap} = y_t^{gap} + e_{i,t}, \tag{4.17}$$

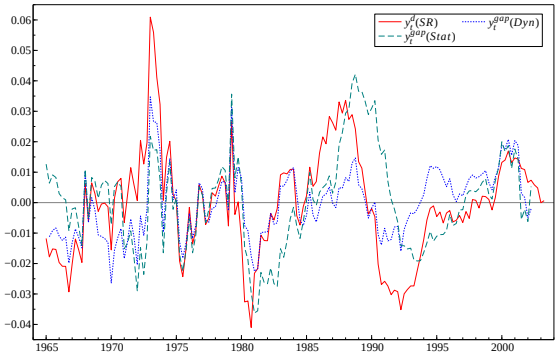

Figure 4.9: The static PF, dynamic PF and Solow residual estimates of the output gap

for $i = 1, ..., 3$, and we postulate $e_{i,t} \sim \mathsf{N}\left[0, \sigma_{e_i}^2\right]$, a classic signal extraction problem exists, where the signal to noise ratio is given by $\frac{\sigma_{\widehat{y}_i}^2}{\sigma_{e_i}^2}$. Principal components analysis should extract the signal relative to the errors, although as this argument requires many strong assumptions, we prefer to work with one estimate of the output gap to maintain interpretability. However, set within this framework, the gap measure we use is given by:

$$\widehat{y}_t^{gap} = 1\left(y_t^{d(SR)}\right) + 0\left(y_t^{gap(Stat)}\right) + 0\left(y_t^{gap(Dyn)}\right). \tag{4.18}$$

PcGets could, in principle, solve for the weights on each measure of the gap within the inflation equation, although collinearity between the measures may be problematic.

4.3 Empirical model of UK inflation

Hendry (2001) argues that there is no single-cause explanation of inflation. In order to test the relevance of all possible causes of inflation we use the *Gets* approach adopted by PcGets. The use of the single-equation framework requires weak exogeneity in the regressors. A failure of weak exogeneity requires a system in which all variables are modelled explicitly, capturing the variety of channels through which correction to the long-run equilibrium takes place. However, the single-equation framework tends to be more robust, and enables a broader analysis, and so we concentrate on this methodology. A VEqCM model is built for the purpose of forecasting in Chapter 5.

The initial model of Δp_t includes three lags of y^d, U^d, e_r, $(c^* - p)$, s, $(ppi - p)$, $(rent - p)$, $(imp - p)$, $(oil - p)$, $(n - p)$, R_l, Δppi, $\Delta rent$, Δimp, Δoil, Δp, Δc^*, $\Delta m4$,

Table 4.1: Indicator variables for the quarterly inflation model

Label	*Values*		*Explanation*
$D_{(73:2,79:3)}$	-1 in 1973q2	$+1$ in 1979q3	Measurement error, VAT increase
$D_{(72:4,74:1)}$	$+1$ in 1972q4	-1 in 1974q1	Oil price shocks
$D_{(84:1,84:2)}$	-1 in 1984q1	$+1$ in 1984q2	Exchange rate fluctuations

Δn, ΔR_s, ΔR_l, Δpw, two lags of the PPP interaction terms $\Delta e_t e_{r,t-j}$ and $\Delta e_t e_{r,t-j}^2$, and an intercept and trend. The PPP interaction terms capture nonlinear exchange rate adjustment asymmetries. These are computed by the difference of the effective exchange rate multiplied by the lagged real exchange rate $[\Delta e_t e_{r,t-j}]$ and the difference of the effective exchange rate multiplied by the lagged real exchange rate squared $\left[\Delta e_t e_{r,t-j}^2\right]$ for $j = 1, 2$. Three blip indicators are included, listed in Table 4.1.

Contemporaneous covariates are excluded from the initial general model in order to reduce the possibility of reverse causation bias in the results. If some of the variables were not pre-determined, a shock may cause a contemporaneous effect on inflation and other $t-$dated variables. For example, an exchange rate shock may impact upon import prices and inflation simultaneously, biasing the results from the inflation model. Furthermore, exclusion of contemporaneous covariates from the model allows us to derive *ex ante* forecasts in Chapter 5. Their inclusion would require known values of the covariates at the forecast horizon, biasing the forecast errors of the partial model downwards in comparison to the VEqCM.

The initial reduction was undertaken by PcGets using the liberal strategy, and a further reduction was imposed by replacing $p_{t-1}, c^*_{t-1}, ppi_{t-1}$ and $e_{r,t-1}$ with the mark-up, π^*_{t-1}, reported in (4.19). We make the assumptions of long-run linear price homogeneity and the adjustment speeds are the same in response to c^*, ppi and e_r.

$$\pi^*_t = p_t - 0.70c^*_t - 0.10ppi_t - 0.20pw_{£,t} + 0.03. \tag{4.19}$$

Imposing this restriction yields $\mathsf{F}_{reduct}(2, 112) = 0.26$ and so the restriction is accepted. Unit labour costs feed through to the GDP deflator with a coefficient of 0.7, which is equivalent to that found in both the production function approaches. Unit labour costs are dominant in determining the price level, and this is consistent with Batini *et al.* (2000),

who find that the labour share (represented by c) is an important leading indicator of UK inflation. The mark-up is adjusted for a zero mean.[6]

The final model is reported in (4.20). The model represents a good fit with a standard error of 0.61%, which is low in view of the turbulence in inflation over the period. All diagnostic and constancy tests are passed. The actual and fitted values are recorded in Figure 4.10, along with the scaled residuals, their correlogram and residual density. The recursive coefficients, 1-step residuals and constancy tests are recorded in Figure 4.11. The recursive graphics show remarkable parameter stability since 1980, other than for a steady gradual decline in the coefficient on the growth rate of producer prices, which is encouraging given its incorrect sign. The 1-step residuals mostly lie within the $\pm 2\sigma$ bands, other than in 1997q1. As the model is relatively stable over time despite many regime changes we can conclude that the implications of the Lucas critique are limited.

$$\begin{aligned}
\Delta p_t &= \underset{(0.029)}{0.064}\Delta imp_{t-1} - \underset{(0.023)}{0.081}\Delta ppi_{t-2} + \underset{(0.033)}{0.223}y^d_{t-1} - \underset{(0.025)}{0.128}U^d_{t-3} \\
&\quad + \underset{(0.176)}{0.483}\Delta e_t e_{r,t-1} + \underset{(0.115)}{0.845}\Delta pw_{t-2} + \underset{(0.038)}{0.105}\Delta^2_4 p_{t-1} - \underset{(0.022)}{0.133}\pi^*_{t-1} \\
&\quad + \underset{(0.005)}{0.049}D_{73:2,79:3} + \underset{(0.005)}{0.020}D_{72:4,74:1} + \underset{(0.004)}{0.016}D_{84:1,84:2} + \underset{(0.002)}{0.006} \qquad (4.20) \\
\mathsf{R}^2 &= 0.858 \quad \widehat{\sigma} = 0.613\% \quad SIC = -9.829 \\
\mathsf{F}_{\mathsf{ar}}(5,109) &= 0.422 \quad \mathsf{F}_{\mathsf{arch}}(4,106) = 0.513 \quad \mathsf{F}_{\mathsf{het}}(22,91) = 0.904 \quad \chi^2_{\mathsf{nd}}(2) = 0.926 \\
\mathsf{F}_{\mathsf{reset}}(1,113) &= 1.311 \quad \mathsf{F}_{\mathsf{Chow}}(20,114) = 0.755 \quad T = 1967q1 - 1998q2.
\end{aligned}$$

4.3.1 Model interpretation

The resulting model contains variables representing most theories of inflation with interpretable signs, other than the growth rate of the producer price index. The results for quarterly post-war inflation are essentially very close to those obtained by Hendry (2001) for annual inflation over the period 1875-1991, suggesting that the modelling approach used does explain inflation well. No evidence of inflation persistence is found, refuting much of the literature that suggests that coefficients of the lagged dependent variable are statistically insignificant from 1. Observed inflation persistence in these models may well

[6] As the prices are indices there is no natural metric for measuring π^*_t.

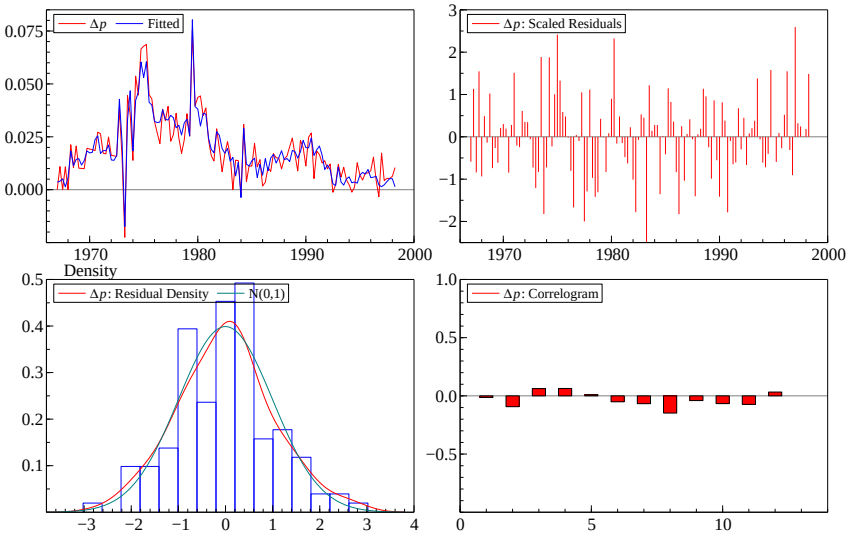

Figure 4.10: Fitted and actual values, residuals, density, and correlogram for the EqCM model of quarterly inflation

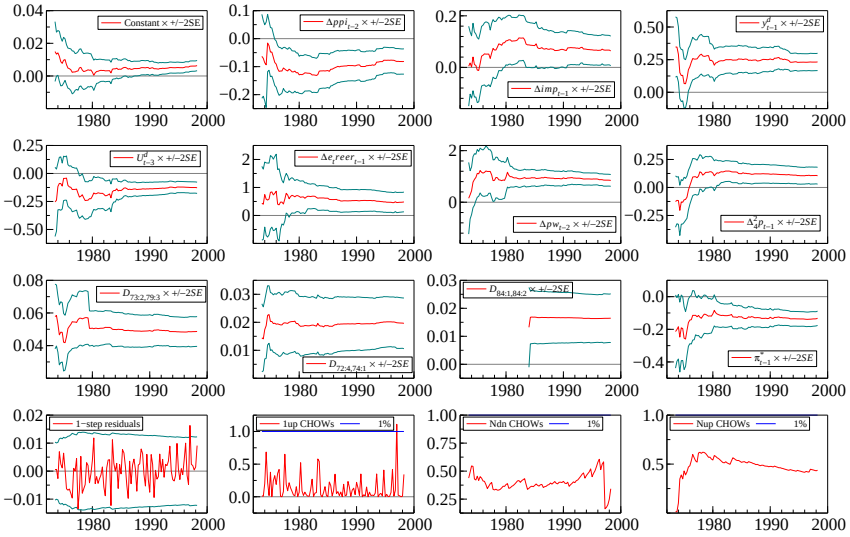

Figure 4.11: Single-equation model of quarterly inflation; recursive coefficients with $\pm 2\sigma$, 1-step residuals and constancy tests

be due to second round effects in explanatory variables that are not modelled. There is a small but significant constant of 0.6%. This is unlikely to be autonomous inflation, and, given the difficulty of estimating the potential levels of the mark-up variables, the most plausible explanation for the non-zero intercept is that $\pi^* \neq 0$ in equilibrium for the units used. Figure 4.3d records π_t^*, which is adjusted for a zero in-sample mean, and this may not directly correspond to the equilibrium over the sample period.

The output gap has a substantial effect upon inflation of 22%, which is highly significant. Replacing the output gap measure by the static production function measure yields a smaller impact of 19%, also entering significantly with one lag, which indicates that the timing of the transmission of a shock from the gap onto inflation is captured consistently. The production function measure does yield a slightly poorer model, with $\widehat{\sigma} = 0.64\%$ and a log-likelihood = 457.79, compared to a log-likelihood of 469.45 for the Solow residual method. We conclude that excess demand for goods and services is a fundamental explanatory variable for the determination of inflation.

Excess demand for unemployment has a significant effect of 13%. To test the robustness of this measure, replacing U^d with the NAIRU calculated using a HP filter led to a significant but smaller impact upon inflation. Unit labour costs enter significantly via the mark-up and the growth rate. The mark-up variable is highly significant, with an effect of 13%. Hence, c^*, e_r and *ppi* are all important determinants of inflation. Both an acceleration of annual inflation term and a nonlinear PPP disequilibrium term are significant. The nonlinear term represents a larger impact from devaluations when there is a greater PPP disequilibrium, particularly with overvaluations.

World inflation has a substantial impact on the UK inflation rate, entering with a two quarter lag. Import prices have a small but significant effect on inflation, and producer prices enter significantly but with the wrong sign. This is likely to be driven by its collinearity with import prices, and the primary input price effect will be driven by the mark-up. Moreover, any mis-measurement in the input price indices could be reflected in opposing signs. The dummies are all highly significant, and as they are capturing one-off shocks or outliers, they should not enter the long-run solution as level shifts. Monetary

terms including the growth rate of broad money and the interest rate are not retained in the model selection process.

The evidence suggests that there is no single cause of inflation. Inflationary pressures arise via many different channels, which can be captured in terms of excess demand or supply pressures in different markets.

4.4 A 'business cycle' factor

An alternative theory proposes that inflation is driven by general 'business cycle' factors that capture all demand and supply pressures, see Stock and Watson (1999a). To examine this theory we test whether it is possible to explain inflation by a few composite factors. Principal components (PC), originating in Hotelling (1933), are statistical techniques used for data reduction by finding linear combinations of the variables that contain most information. We shall estimate a composite measure of the business cycle based on PC analysis in order to assess whether information is lost by explaining inflation by general business cycle factors, as opposed to excess demand and supply pressures from all sectors of the economy.

4.4.1 Principal component analysis

Muirhead (1982) and Anderson (1994) provide outlines of the PC approach. Assume a random $m \times 1$ vector $\mathbf{X}$ has a normal distribution, $\mathsf{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, and let $\lambda_1, \geq \lambda_2 \geq \cdots \geq \lambda_m$ (> 0) be the latent roots of $\boldsymbol{\Sigma}$. The $m \times m$ orthogonal matrix of eigenvectors, $\mathbf{H} = [\mathbf{h}_1 ... \mathbf{h}_m]$, implies:

$$\mathbf{H}'\boldsymbol{\Sigma}\mathbf{H} = \boldsymbol{\Lambda} = \mathsf{diag}(\lambda_1, ..., \lambda_m). \tag{4.21}$$

$\mathbf{U}$ is defined as:

$$\mathbf{U} = \mathbf{H}'\mathbf{X} = (U_1, ..., U_m)', \tag{4.22}$$

where $\mathsf{cov}[\mathbf{U}] = \boldsymbol{\Lambda}$, and hence, $U_1, ..., U_m$ are uncorrelated and $\mathsf{V}[U_i] = \lambda_i, i = 1, ..., m$. The components $U_1, ..., U_m$ of $\mathbf{U}$ are the PCs of $\mathbf{X}$, and the first PC is given as $U_1 = \mathbf{h}_1'\mathbf{X}$ with variance λ_1. This is the normalised linear combination of the components of $\mathbf{X}$ with the largest possible variance. The second PC will then account for the maximum of the

remaining variance, and all the components are derived in this manner. The method serves to combine all variables into a composite variable that reflects the maximum possible proportion of the total variation in the set.

The PCs are determined under the condition that they are orthogonal. If we define an arbitrary linear function as $\boldsymbol{\alpha}'\mathbf{X}$ with $\mathsf{V}[\boldsymbol{\alpha}'\mathbf{X}] = \boldsymbol{\alpha}'\boldsymbol{\Sigma}\boldsymbol{\alpha}$, the condition that $\boldsymbol{\alpha}'\mathbf{X}$ is uncorrelated with the ith PC, U_i, is:

$$0 = \mathsf{cov}\left[\boldsymbol{\alpha}'\mathbf{X}, \mathbf{h}_i'\mathbf{X}\right] = \boldsymbol{\alpha}'\boldsymbol{\Sigma}\mathbf{h}_i = \lambda_i\boldsymbol{\alpha}'\mathbf{h}_i, \tag{4.23}$$

as $\boldsymbol{\Sigma}\mathbf{h}_i = \lambda_i\mathbf{h}_i$, so $\boldsymbol{\alpha}$ must be orthogonal to $\mathbf{h}_i$. Two measures that explain the variability in $\mathbf{X}$ are tr$\boldsymbol{\Sigma}$ and det$\boldsymbol{\Sigma}$, where:

$$\mathrm{tr}\boldsymbol{\Sigma} = \mathrm{tr}\mathbf{H}'\boldsymbol{\Sigma}\mathbf{H} = \mathrm{tr}\boldsymbol{\Lambda} = \sum_{i=1}^{m}\lambda_i, \tag{4.24}$$

$$\det\boldsymbol{\Sigma} = \det\mathbf{H}'\boldsymbol{\Sigma}\mathbf{H} = \det\boldsymbol{\Lambda} = \prod_{i=1}^{m}\lambda_i. \tag{4.25}$$

For the sample PCs, suppose $X_1, ..., X_N$ is a random sample of size $N = n + 1$ on $\mathbf{X}$. We can define the sample covariance matrix, $\mathbf{S}$, by:

$$\mathbf{A} = n\mathbf{S} = \sum_{i=1}^{N}\left(\mathbf{X}_i - \overline{\mathbf{X}}\right)\left(\mathbf{X}_i - \overline{\mathbf{X}}\right)'. \tag{4.26}$$

The latent roots of $\mathbf{S}$ (labelled $l_1, ..., l_m$) are estimates of the latent roots $\lambda_1, \geq \cdots \geq \lambda_m$ of $\boldsymbol{\Sigma}$. Defining the matrix of normalised eigenvectors, $\mathbf{Q} = [\mathbf{q}_1 ... \mathbf{q}_m]$ such that:

$$\mathbf{Q}'\mathbf{S}\mathbf{Q} = \mathbf{L} = \mathsf{diag}\left(l_1, ..., l_m\right), \tag{4.27}$$

we can estimate the eigenvector $\mathbf{h}_i$ by the sample. The sample PCs are given as $\widehat{U}_1, ..., \widehat{U}_m$ of $\widehat{\mathbf{U}} = \mathbf{Q}'\mathbf{X}$.

While the decision as to how many factors to extract is arbitrary (because of the lack of interpretation that can be given to the factors), two methods often used include the Scree test (Cattell, 1966) and the Kaiser (1960) criterion. The Kaiser criterion suggests retaining factors with eigenvalues greater than one, as the factor is only then extracting at least as much information as the original variable. The Scree test suggests plotting the eigenvalues and determining the number of factors occularly based on where the plot

smoothes out horizontally.

4.4.2 Modelling inflation with principal components

We develop two PC models of inflation. First, a model of inflation based on the PCs and lagged dependent variable is developed, and its performance is compared with the inflation model in (4.20). Second, the dominant congruent inflation model (4.20) is augmented by PCs and selection is then undertaken, allowing both individual variables and PCs to explain inflation. To avoid perfect collinearity between the PCs and individual regressors if commencing from the full information set, the GUM will consist of only those variables retained in (4.20), along with the first seven PCs. The theoretical underpinnings of the PC approach lie in the decision as to which variables to include. Stock and Watson (1998) adopt a very general approach, whereby they include 216 variables in the analysis. As the main aim of our analysis is to detect a general structure in the combined variables, a smaller subset of data is used. This ensures the results are comparable with those of section 4.3, and avoids cluttering with irrelevant variables that may pick up spurious correlations.

Table 4.2 reports the estimated eigenvalues for the first seven PCs based on the variables: ppi, wpi, c^*, oil, $rent$, n, imp, U_r, R_s, R_l, $m4$, e_r, $asset$, U^d and y^d. Two additional variables are included compared to the GUM in section 4.3; the FTSE all share index ($asset$), and the manufacturing wholesale price index (wpi). Both levels and first differences are included in order to detect trend and cycle components.[7] The variables are normalised by $x_i^* = \frac{(x_i - \mu_{x_i})}{\sigma_{x_i}}$ to avoid the series with the greatest amplitude in cycle exerting too much pressure on the PC. The Kaiser criterion and Scree test suggest that six or seven components should be retained.

Figure 4.12 records the first four PCs, recorded with the price level for PC_1 and quarterly inflation for PC_2, PC_3 and PC_4. The first component closely follows the trend in the price level, although it is much more volatile, the second component matches inflation reasonably well (ρ= 0.67), and the third and fourth components are picking up

[7] The first difference of both housing rent and national debt are excluded from the principal component analysis as the levels of these variables are very smooth, and hence the low variance of the differences adversely biases the components. Furthermore, p_{t-1} and Δp_{t-1} are excluded to avoid biasing the results.

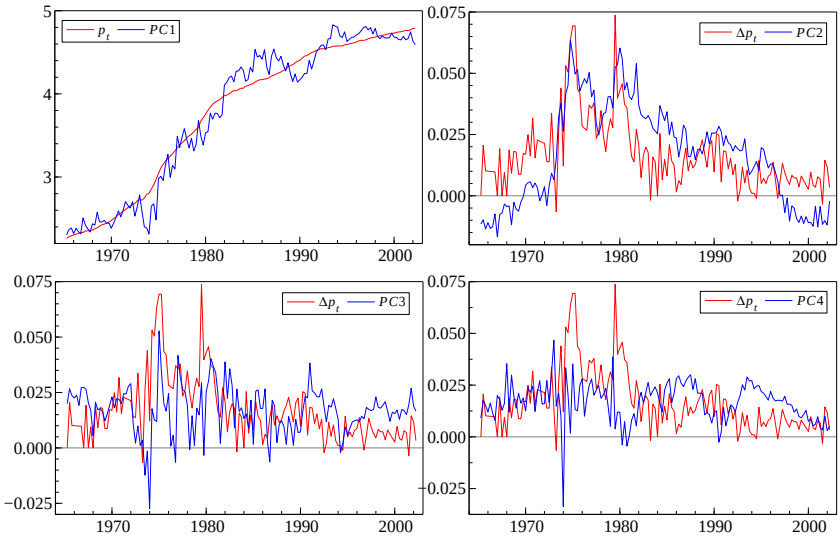

Figure 4.12: Leading four principal components for quarterly inflation

Table 4.2: Estimated eigenvalues: first seven principal components of inflation.

	Eigenvalues	*Cumulative %*
$PC1$	8.994	32.12
$PC2$	4.250	47.30
$PC3$	2.877	57.58
$PC4$	2.647	67.03
$PC5$	1.718	73.16
$PC6$	1.374	78.07
$PC7$	1.216	82.42

Notes: cumulative % reports the cumulative percentage of variation explained by the principal components.

innovations in the data.

The PcGets conservative strategy is used to estimate an inflation model based on the principal components. The GUM consists of 33 regressors, including four lags of the first seven PCs, imposing the restriction that the first component enters in differences, along with lags of the dependent variable, an intercept, and three indicator dummies for outliers. Contemporaneous PCs are excluded to ensure comparability with (4.20). The resulting model is reported in (4.28).

$$\begin{aligned}
\Delta p_t &= \underset{(0.001)}{0.004} + \underset{(0.065)}{0.240}\Delta p_{t-1} + \underset{(0.064)}{0.508}\Delta p_{t-3} + \underset{(0.238)}{0.778}\Delta PC_{1,t-1} - \underset{(0.050)}{0.198}PC_{3,t-1} \\
&\quad - \underset{(0.054)}{0.248}PC_{3,t-3} - \underset{(0.055)}{0.154}PC_{4,t-3} - \underset{(0.009)}{0.045}I_{73:2} + \underset{(0.009)}{0.036}I_{75:2} + \underset{(0.009)}{0.050}I_{79:3} \\
\mathsf{R}^2 &= 0.729 \quad \widehat{\sigma} = 0.841\% \quad SIC = -6.418 \quad \mathsf{F}_{\mathsf{ar}}(5,111) = 1.599 \qquad (4.28) \\
\chi^2_{\mathsf{nd}}(2) &= 1.860 \quad \mathsf{F}_{\mathsf{arch}}(4,108) = 1.094 \quad \mathsf{F}_{\mathsf{reset}}(1,115) = 3.448 \\
\mathsf{F}_{\mathsf{het}}(15,100) &= 0.805 \quad \mathsf{F}_{\mathsf{Chow}}(20,116) = 0.457 \quad T = 1967q1 - 1998q2.
\end{aligned}$$

The model fit is poorer than the inflation model in (4.20), with a residual standard error of 0.84% compared to 0.61%. Hence, the PCs do not explain inflation as well as the individual regressors. One concern with principal component models is that as the PC weights are fixed throughout the sample, any structural change in the relationships between the individual regressors will be forced to manifest itself through parameter non-constancy. The large coefficients on the lagged dependent variables suggest that the LDV is capturing shifts in the underlying inflation process as opposed to the PCs. The second PC is not retained, but as the lagged dependent variables are highly significant, this is consistent with our view that $PC_{2,t}$ is proxying inflation. While the composite measure does model inflation reasonably well, the inability to interpret the model implies that the model is of limited value to policy-makers, and we conclude that modelling inflation with individual explanatory variables, selected within a *Gets* framework, provides the more coherent and sagacious strategy.

A further experiment undertaken to test the explanatory power of the PCs is to include the business cycle factors in the inflation model derived in (4.20), and then select the dominant model using PcGets. This allows both the individual variables and business cycle factors to explain inflation, with the selection algorithm determining which variables have the most explanatory power. The GUM consists of the union of the twelve variables in (4.20) and four lags of the first seven principal components, excluding contemporaneous

regressors. A conservative strategy is used to select the resulting model:

$$\begin{aligned}
\Delta p_t &= \underset{(0.030)}{0.121}\Delta imp_{t-1} - \underset{(0.023)}{0.098}\Delta ppi_{t-1} + \underset{(0.031)}{0.211}y^d_{t-1} + \underset{(0.168)}{0.404}\Delta e_t e_{r,t-1} \\
&\quad + \underset{(0.049)}{1.336}\Delta pw_{t-2} + \underset{(0.036)}{0.097}\Delta^2_4 p_{t-1} - \underset{(0.025)}{0.154}\pi^*_{t-1} - \underset{(0.044)}{0.184}PC_{2,t-1} \\
&\quad + \underset{(0.038)}{0.109}\Delta PC_{5,t-1} - \underset{(0.047)}{0.198}PC_{6,t-2} + \underset{(0.004)}{0.053}D_{73:2,79:3} \\
&\quad + \underset{(0.004)}{0.017}D_{72:4,74:1} + \underset{(0.004)}{0.015}D_{84:1,84:2} \qquad (4.29) \\
\widehat{\sigma} &= 0.585\% \quad SIC = -9.891 \quad \mathsf{F_{ar}}(5,108) = 1.021 \\
\chi^2_{\mathsf{nd}}(2) &= 0.885 \quad \mathsf{F_{arch}}(4,105) = 1.525 \quad \mathsf{F_{reset}}(1,112) = 0.589 \\
\mathsf{F_{het}}(26,86) &= 0.524 \quad \mathsf{F_{Chow}}(20,113) = 1.633 \quad T = 1967q1 - 1998q2.
\end{aligned}$$

The model passes all diagnostics and is almost equivalent to (4.20), with a residual standard error of 0.59% compared to 0.61%. The model retains all but one of the original regressors, excluding excess demand for unemployment. Given the significance of U^d in (4.20), high correlations between U^d and the PCs must explain this result. The model, with thirteen parameters, only retains four principal components (the restriction that $PC_{5,t-1}$ and $PC_{5,t-2}$ are reparameterised to enter as $\Delta PC_{5,t-1}$ is accepted and therefore implemented), although selection using the liberal strategy would increase this to nine. The main determinants of inflation mostly have similar coefficients to (4.20). Two notable exceptions are Δpw_{t-2} and Δimp_{t-1}. The coefficient on import prices has almost doubled, and world inflation has a coefficient of 1.3, a consequence of the substantial reduction in the coefficient standard error for Δpw_{t-2}. It is difficult to interpret the model coefficients because the principal components will contain combinations of all regressors, generating collinearity between the PCs and individual regressors, although in principle, the factor loadings could be used to calculate the overall effect for each relevant variable. The PCs do not negate the dummies, suggesting that these are modelling effects that are not captured by the economic variables included.

The evidence does suggest that while this data reduction method does capture useful information, it cannot substitute well-specified reduced form equations that attempt to model all significant theories of inflation. The single-cause explanation of inflation, in

this case represented by what we term general 'business cycle characteristics', is again refuted. The problems of a lack of interpretability and non-robustness to changes in the information set considerably hinder the use of principal component methods. The similarity between (4.20) and (4.29) provides strong evidence for undertaking selection on individual variables as opposed to principal components. Furthermore, given the techniques to handle more variables than observations discussed in Chapter 2, the need for PCs as data reduction methods is abated.

4.5 Conclusion

The chapter develops a dominant congruent model of quarterly UK inflation from 1967q1-1998q2 that encompasses all relevant theories of inflation and is a good representation of the data. The central contribution of the chapter is the use of general-to-specific methodology to select an in-sample model. The *Gets* methodology is based on data informing the model specification, as opposed to the testing of theoretical priors on the data. This enables all possible theories to be tested within the same framework, increasing the probability of locating the underlying LDGP. However, theory has played a fundamental role in the specification of the inflation equation through the choice of variables entering the GUM. All the explanatory variables are based on theoretical models, including the specification of the excess demand variables for both the goods and labour markets, the mark-up, and the nonlinear PPP interaction term. The use of an automatic model selection procedure such as PcGets enables many more paths to be explored and this avoids potential path dependence. The initial GUM contains 75 potential explanatory variables, and it would be almost impossible to undertake a full search by hand. The algorithm not only ensures a rigorous search procedure, but it also undertakes specification tests at every stage to ensure the final model is well-specified. Hence, all extant theories of inflation can be tested within a systematic and coherent framework, enabling us to make clear statements about the determinants of inflation.

Three measures of goods market pressures, or the output gap, are developed. A production function method is initially undertaken in a static and cointegrating framework. The residual estimate of TFP accords with our priors regarding this latent variable, and

the lack of cyclicality suggests that efforts to correct for labour hoarding and capacity utilisation are successful. Given the presence of substantial and systematic measurement error in the capital stock, potential output is then modelled as the long-run solution to a dynamic model with a time-varying intercept that proxies TFP. The dynamic model attributes more of the fluctuations in output to changes in potential output, resulting in a smaller gap. Finally, the gap is measured as the Solow residual from a measure of capacity that uses a split time trend to capture changes in productivity. Removing the stochastic TFP component results in a larger gap, and this measure is used in the inflation model.

Key determinants of inflation are identified as the mark-up, capturing impacts from unit labour costs, producer prices, exchange rates and foreign prices, excess demands for both goods and services and labour, foreign inflation and import prices. The change in domestic inflation and a nonlinear term capturing exchange rate adjustment asymmetries are also significant. The results concur with those for annual UK inflation over 1875-1991 derived in Hendry (2001), and therefore suggest that the modelling approach used does explain inflation well.

The *Gets* framework enables us to test the hypothesis that inflation is driven by a general 'business cycle' component. Principal components are derived from the information set used in the inflation model and a model of inflation is developed using PcGets. A substantial increase in the residual standard error of 23% indicates that a general 'business cycle' explanation of inflation is not sufficient, and the inability to interpret the results further hampers the modelling approach. Pitting both potential explanations (individual variables and principal components) against each other within the *Gets* framework results in the selection of almost all individual determinants of inflation and very few principal components. Hence, the information gained by including the principal components is relatively minimal, and we conclude that the channels by which various sectoral pressures feed through to inflation need to be modelled individually.

4.A Appendix

Table 4.3: Data Appendix

Variable	*Description*	*Source*	*Code*
Y_t	Gross domestic product: chained volume measure	NS	ABMI
P_t	Gross domestic product (expenditure) at market prices deflator	NS	YBGB
PPI_t	PPI manufacturing input - raw materials	DS	UKOPP029F
IMP_t	Import price index	DS	UKIMPPRCF
$RENT_t$	Actual rentals for housing* + imputed rentals for housing*, £m.	NS	ADFT+ADFU
C_t	Unit labour cost index for the whole economy	NS	LNNL
PW_t	OECD consumer price index	DS	OCICP009F
OIL_t	World market price of crude petroleum	DS	WD176AAZA
N_t	Public sector net debt, £m.	NS	BKQK
$M4_t$	Nominal broad money stock (end period), £m.	NS	AUYN
$R_{s,t}$	Three-month treasury bill rate	DS	UKGBILL3
$R_{l,t}$	Yield on 20-year gilts	DS	UKGBOND
E_t	Sterling effective exchange rate index	DS	UKXTW..NF
GVA_t	Gross value added at basic prices: chained volume measure	NS	ABMM
$ASSET_t$	FTSE all share index/$(GVA \times P)$	DS	UKSHRPRCF
WPI_t	Wholesale price index of materials and fuel in manufacturing	DS	UKPPIMMNF
W_t	Total compensation of employees, current price, £m.	NS	DTWM
Z_t	Total gross operating surplus*, current price, £m.	NS	ABNF
J_t	Net capital stock for whole economy exc. dwellings sector, £m.	BoE	–
I_t	Total gross fixed capital formation, constant price, £m.	NS	NPQT
ΔINV_t	Changes in inventories, constant price	NS	CAFU
OH_t	Weekly overtime hours per operative $\times \frac{\text{fraction on overtime}}{\text{average hours}}$	EPG, DEG, EG, LMT	
$U_{c,t}$	Capacity utilisation based on CBI index: % working below capacity, see derivation in appendix	DS	UKCBICAB
NH_t	Ave. actual weekly hours of work (all workers: main & 2nd job)	NS	YBUV from 1992
		Pre-1992:	EPG, DEG, EG
$WPOP_t$	Population aged 16-59/64, '000s	NS	YBTF from 1992
		Pre-1992:	EPG, DEG, EG
EMP_t	Total number in employment, aged 16+, '000s	NS	MGRZ from 1992
		Pre-1992:	EPG, DEG, EG
$INACT_{r,t}$	Economically inactive population / population, 16+, '000s	NS	MGSI / MGSL
$EMP_{r,t}$	$EMP_t/WPOP_t$		
U_t	$WPOP_t - EMP_t$		
$U_{r,t}$	$U_t/WPOP_t$		
$PART_{r,t}$	$1 - INACT_{r,t}$		
L_t	$WPOP_t \times EMP_{r,t} \times PART_{r,t}$		

Notes: All data are seasonally adjusted. * denotes variable has been seasonally adjusted using X11.
Sources include [NS] National Statistics database; [IFS] International Financial Statistics Database; [BoE] Bank of England; [DS] Datastream; [EPG] Employment and Productivity Gazette, pre-1971; [DEG] Department of Employment Gazette, 1971-79; [EG] Employment Gazette, 1980-1995; [LMT] Labour Market Trends, 1996-present.

Calculation of a capacity utilisation index

The CBI industrial trends survey reports the response of firms in the manufacturing sector to the question: "Is your present level of output below capacity?" (defined as a satisfactory or full rate of operation). Following Muellbauer (1984), we can define the

proportionate deviation of capacity utilisation U_c by:

$$-U_c = \ln Y(max) - \ln Y. \tag{4.30}$$

If different firms have the same view regarding satisfactory levels of operation, we can define Z as:

$$Z = \ln Y(max) - \ln Y(sat). \tag{4.31}$$

Assuming a distribution of utilisation across firms, measured by $\ln Y(max) - \ln Y$, which shifts through time with a limit fixed at zero, we can calculate the proportion of firms operating below the usual level of capacity π which is observed, and link this with the unobserved mean of the distribution $\mathsf{E}(-U_c)$. If the distribution of capacity is lognormal:

$$\ln(-U_c) \sim \mathsf{N}\left[\mu, \sigma_{U_c}^2\right], \tag{4.32}$$

then

$$\pi = 1 - \Phi\left(\frac{\ln Z - \mu}{\sigma_{U_c}}\right). \tag{4.33}$$

Therefore,

$$\mathsf{E}\left[\ln(-U_c)\right] = \mu = \ln Z - \sigma_{U_c}\Phi^{-1}(1-\pi), \tag{4.34}$$

where $\Phi(.)$ is the standard normal distribution. Due to the normality assumption, we can derive $\mathsf{E}(-U_c)$ from $\mathsf{E}\left[\ln(-U_c)\right]$:

$$\mathsf{E}(-U_c) = \exp\frac{1}{2}\sigma_{U_c}^2 \exp\mu = \beta \exp\left[-\sigma_{U_c}\Phi^{-1}(1-\pi)\right], \tag{4.35}$$

where $\beta = Z\exp\frac{1}{2}\sigma_{U_c}^2$. Amemiya (1981) suggests approximating by a logistic distribution:

$$\Phi(x) \approx \frac{\exp(1.6x)}{1+\exp(1.6x)} \sim \left[0, \frac{1.6^2}{\pi^2/3}\right]. \tag{4.36}$$

to deliver:

$$\mathsf{E}(-U_c) = \beta\left(\frac{\pi}{1-\pi}\right)^{\frac{\sigma_{U_c}}{1.6}}. \tag{4.37}$$

Muellbauer (1984) recommends empirical magnitudes of $\sigma_{U_c} = 0.64$ and $Z = 0.09$, which suggests that full capacity is approximately 91% of the physical maximum. These values were found by estimating a production function with $\left(\frac{\sigma_{U_c}}{1.6}\right)$ ranging from 0.2 to 0.6.

Chapter 5

Forecasting UK Inflation: Empirical Evidence on Robust Forecasting Devices

5.1 Introduction

Systematic mis-forecasting of economic outcomes has led to extensive research on economic forecasting, culminating in a new theory of forecasting developed by Clements and Hendry (1998b, 1999, 2002a). This theory refutes the common assumption that the econometric model coincides with the DGP in a stationary world. The forecast error taxonomy developed allows for a mis-specified model with measurement error in the data, within a non-stationary world that is subject to structural breaks. This is more representative of the conditions in which forecasts of UK inflation are derived. This chapter examines the forecast performance of both econometric models and forecasting rules for UK annual and quarterly inflation, using the Clements and Hendry forecast theory to explain the subsequent results.

We utilise the model of UK inflation built in Chapter 4 to forecast inflation. However, concerns over the weak exogeneity assumptions made in developing the single-equation model in order to sustain the conditioning on the non-modelled variables, and the subsequent requirement of strong exogeneity needed to forecast more than 1-step ahead (see Engle *et al.*, 1983) suggests that endogenizing all variables in the form of a VAR may yield preferable results. Furthermore, the VAR framework enables dynamic forecasts to be computed. While, in principle, the size of the VAR should not be prohibited by the sample size available, as the techniques developed to handle more variables than observations (discussed in Chapter 2) are also applicable to the VAR framework, in practice the current software is more restrictive. Multi-stage estimation using subsets of regressors would enable a VAR in the 23 exogenous variables conditioned on in the single-equation framework to be constructed. Indeed, one could think of the single-equation EqCM developed in Chapter 4 as one equation of a VEqCM, as it only uses lagged variables.

However, we build a VAR model of inflation outside of the *Gets* framework in which only five endogenous variables are modelled, and this abstracts from the degrees of freedom problems associated with a more general VAR. A congruent VEqCM model of inflation is built over the same in-sample period, and both the single-equation and vector equilibrium correction models are used to forecast inflation over the period 1998-2003. Furthermore, various robust transformations of the models are undertaken in an attempt to 'robustify' forecasts to structural breaks, and we consider various differencing devices, rapid updating tools, and forecast pooling, to overcome the problem of forecast failure.

We tentatively conclude that robust forecasting devices do prove useful in forecasting macroeconomic time-series, and they often outperform the dominant congruent in-sample model, both when there are structural breaks in the data and when the underlying process appears to be stable but with probable breaks in the explanatory variables. We also conclude that increasing the information set does lead to improvements in forecasting performance, suggesting that disaggregation can yield benefits. Finally, we observe that much of the forecast error in the structural models is driven by the deterministic terms. Breaks in the mean of the cointegrating vector or the growth rate of the system will cause forecast 'failure', and results show how sensitive forecasts are to errors in these terms.

The structure of the chapter is as follows. Section 5.2 develops a multi-equation model of inflation based on various input prices and conditioning on excess demand. An analysis of measurement error in the output gap is undertaken to justify exclusion of the cumulated gap from the cointegrating space. Cointegration analysis is conducted on the unrestricted VAR allowing for a partial system, and a parsimonious VEqCM model is derived. Section 5.3 examines the 1-step and 4-step forecast performance of the single-equation and vector EqCMs of UK inflation against robust forecasting devices, ranking the models in an attempt to predict which models should forecast well. Section 5.4 addresses the theory of predictability, presenting evidence for this by deriving two models of annual inflation, including a lower frequency model using annual analogues and a higher frequency model using quarterly data. The forecasting performance of these annual inflation models are assessed against robust forecasting devices in Section

5.5. Section 5.6 concludes. This chapter draws on the data outlined in Appendix 4.A.

5.2 Model of quarterly inflation

Our single-equation analysis of quarterly inflation led to a mark-up model with an equilibrium correction mechanism involving the real exchange rate, world prices in sterling, unit labour costs and producer prices. We now move to the vector-equation framework by examining a five variable system of inflation based on input prices and conditioning on excess demand. The dynamic system framework allows us to test the restriction of weak exogeneity that is automatically imposed in a single-equation framework, especially with contemporaneous variables. However, we retain the restriction of no contemporaneous covariates to ensure the results are comparable with the single-equation model. By conditioning on excess demand we require the output gap to be weakly exogenous, which is tested for. Furthermore, monetary variables are not included as they are found to be insignificant in the single-equation analysis.

5.2.1 Conditioning on the output gap

Inclusion of the output gap, a stationary explanatory variable, results in a nuisance parameter. Rahbek and Mosconi (1999) show that the cumulated output gap should be included in the equilibrium correction mechanism, enabling the critical values for the trace test computed in Harbo, Johansen, Nielsen and Rahbek (1998) to be used. Without the cumulated gap in the cointegration space, the asymptotic distribution of the trace test is affected by nuisance parameters. However, this analysis shall proceed by just including the lagged output gap outside the cointegration space, and so the trace test statistics are used as an indication of the rank of the cointegrating vector and caution should be applied as the critical values are not exact. This avoids cumulating the I(0) measurement error in the output gap to an I(1) measurement error.

The standard errors associated with the output gap estimates are large, as is demonstrated by the disparity between the three gap estimates derived in Chapter 4 and recorded in Figure 4.9, reproduced in Figure 5.1a for clarity. Hence, the cumulated gaps, recorded in Figure 5.1, panels b–d, for the Solow residual, static production function and

dynamic production function respectively, differ substantially. The cumulated gaps are recorded on the same axes for comparison. The initial conditions have a large impact on the cumulated gap, and yet it is extremely difficult to estimate the origin accurately. The resulting cumulated gap estimates are very different if an alternative date is selected as the origin, and while the gap should be in equilibrium at $t = 0$ (i.e., a zero gap estimate), even conditioning on this still results in different cumulated gap estimates. The dynamic production function gap is permanently negative, which is implausible. The cumulated Solow residual gap does correspond to the booms and recessions observed historically, but the magnitudes are difficult to justify, particularly with regard to the late 1980s boom and subsequent recession. The vastly differing cumulated gap estimates support our decision to exclude the variable from the cointegration analysis.

Results of the augmented Dickey-Fuller (ADF) test for the cumulated output gaps are reported in Table 5.1. The ADF tests are based on a restricted constant and no trend, as Nielsen and Rahbek (2000) argue that a restricted constant is preferable to an unrestricted constant. Inclusion of an unrestricted constant results in a linear trend plus random walk under the null, whereas the restricted constant delivers a random walk under the null. The test statistic is the F-test of $\mathsf{H}_0 : \mu = \gamma = 0$, for:

$$\Delta y_t = \mu + \gamma y_{t-1} + \epsilon_t, \tag{5.1}$$

where $\epsilon_t \sim \mathsf{IN}\left[0, \sigma_\epsilon^2\right]$ for $t = 1, ..., T$, and $\gamma = \beta - 1$ for the regression $y_t = \mu + \beta y_{t-1} + \epsilon_t$. The critical values are taken from Patterson (2000, Table 6.5), for $T = 150$. The dynamic production function gap is I(0), such that the cumulated gap is I(1). Results for the Solow residual gap and the static production function gap are less clear. There is evidence that the static PF gap is I(1), resulting in an I(2) cumulated gap. The Solow residual gap is marginal; we conclude that it is I(0) at the 5% significance level but not at the 1% level. If the cumulated gap is I(1), but is close to being I(2), its inclusion in the cointegrating space may generate a cointegrating vector that is close to I(1). Caveats regarding the power of the ADF test apply.

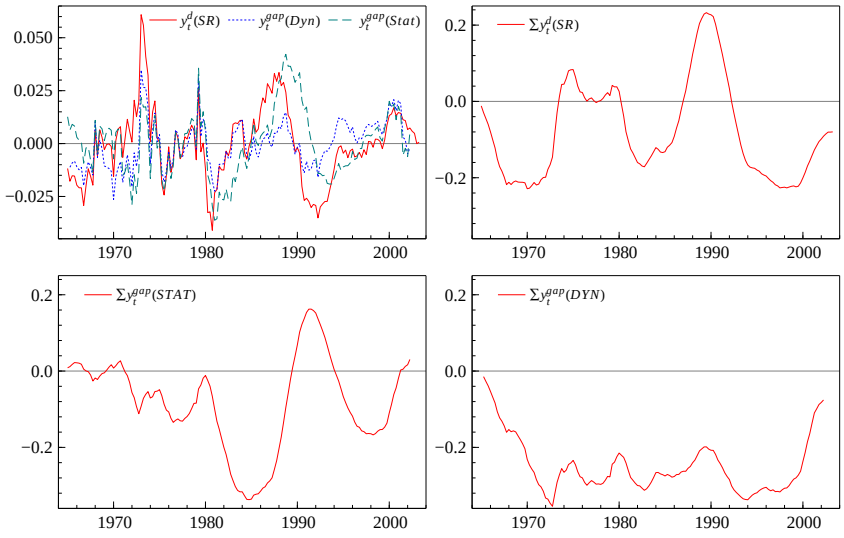

Figure 5.1: Estimates of the output gap and corresponding cumulated gaps

Table 5.1: Augmented Dickey-Fuller tests for the cumulated output gap

Null	*Solow residual gap*			*Static PF gap*			*Dynamic PF gap*		
	$\widehat{\beta}$	$\widehat{\tau}$	lag	$\widehat{\beta}$	$\widehat{\tau}$	lag	$\widehat{\beta}$	$\widehat{\tau}$	lag
I(1)	0.991	0.758	0	0.992	0.493	0	0.985	4.081	1
I(2)	0.869	6.623*	1	0.872	4.640	2	0.759	15.19**	1
I(3)	-0.172	208.3**	1	-0.220	42.25**	3	-0.484	124.8**	1

Notes: ADF statistics based on a restricted constant and no trend. $\mathsf{H}_0 : \mu = \gamma = 0$, with test statistic: $\widehat{\tau} \sim \mathsf{DF}_\tau$. Critical values are 6.73 at 5% significance and 4.72 at 1% significance, for $T = 150$. * and ** denote rejection at the 5% and 1% significance levels.

5.2.2 Measurement errors in the cumulated output gap

One explanation for the empirical results suggesting the cumulated gap is close to I(2) is due to pernicious measurement errors in the estimates of the output gap. All three estimates comprise of labour and capital inputs, which are difficult to measure accurately, not just as is well known with regard to the capital stock, where scrapping and technological change effects are important, but also for labour which should be measured by 'human capital hours'. Hence, we shall assess the impact of measurement error on the cumulated gap, undertaking a Monte Carlo experiment to derive the null rejection frequencies of the ADF test statistic on the cumulated gap. For reasonable levels of persistence we find that it is often difficult to reject the null that the cumulated gap is I(2).

We proxy the output gap with an autoregressive DGP that is contaminated with

measurement error from both capital and labour input:

$$y_t = \rho y_{t-1} + u_{1,t} + u_{2,t} + \epsilon_t, \quad 0 < \rho < 1, \tag{5.2}$$

where $\epsilon_t \sim \mathsf{IN}\left[0, \sigma_\epsilon^2\right]$ for $t = 1, ..., T$, and the initial condition is given by $y_0 = 0$. $u_{1,t}$ corresponds to measurement error in labour input, such that $u_{1,t} \sim \mathsf{IN}\left[0, \sigma_{11}\right]$. $u_{2,t}$ corresponds to measurement error in the capital stock, which exhibits some persistence but is stationary:

$$u_{2,t} = \lambda u_{2,t-1} + v_t, \quad v_t \sim \mathsf{IN}\left[0, \sigma_{22}\right], \quad 0 < \lambda < 1, \tag{5.3}$$

with $\mathsf{cov}\left[u_{1,t}, u_{2,t}\right] = 0$. The measurement error, $u_{1,t} + u_{2,t}$, is unobserved, and therefore is compounded into the error term, such that we observe:

$$y_t = \widehat{\rho} y_{t-1} + \widehat{\epsilon}_t^{\,*} \tag{5.4}$$

where $\epsilon_t^* = u_{1,t} + u_{2,t} + \epsilon_t$. Empirical estimates of $\widehat{\rho}$ based on the three output gap measures suggest $\widehat{\rho} \approx 0.8$.[1]

Given our estimate of $\widehat{\rho}$, and postulating a sensible value of $\lambda = 0.75$, implies we can obtain an estimate of ρ to parameterise the Monte Carlo DGP. Observe that:

$$\widehat{\rho} = \frac{\sum_{t=1}^{T} \left(y_t y_{t-1}\right)}{\sum_{t=1}^{T} \left(y_{t-1}^2\right)}. \tag{5.5}$$

Using:

$$\frac{1}{T} \sum_{t=1}^{T} \left(y_t u_{2,t}\right) \xrightarrow{p} \mathsf{E}\left(y_t u_{2,t}\right) = \frac{\sigma_{22}}{1 - \lambda^2}, \tag{5.6}$$

we can obtain:

$$\begin{aligned} \frac{1}{T} \sum_{t=1}^{T} \left(y_t y_{t-1}\right) \xrightarrow{p} \mathsf{E}\left(y_t y_{t-1}\right) &= \rho^2 \sigma_y^2 + \lambda \mathsf{E}\left(y_{t-1} u_{2,t-1}\right) \\ &= \rho^2 \sigma_y^2 + \frac{\lambda \sigma_{22}}{1 - \lambda^2}, \end{aligned} \tag{5.7}$$

where $\sigma_y^2 = \mathsf{E}\left(y_t^2\right)$. Also:

$$\frac{1}{T} \sum_{t=1}^{T} \left(y_t^2\right) \xrightarrow{p} \mathsf{E}\left(y_t^2\right) = \rho^2 \sigma_y^2 + \sigma_{11} + \frac{\sigma_{22}}{1 - \lambda^2} + \sigma_\epsilon^2 + \frac{2\lambda\rho\sigma_{22}}{1 - \lambda^2}, \tag{5.8}$$

[1] $\widehat{\rho} = 0.85$ for the Solow residual gap, 0.84 for the static production function gap, and 0.72 for the dynamic production function gap.

to deliver:

$$\sigma_y^2 = \frac{\sigma_{11} + \left(\frac{1+2\lambda\rho}{1-\lambda^2}\right)\sigma_{22} + \sigma_\epsilon^2}{1-\rho^2}. \tag{5.9}$$

Inserting (5.7) and (5.9) into (5.5) results in:

$$\text{plim}\widehat{\rho} = \rho + \frac{\lambda\sigma_{22}\left(1-\rho^2\right)}{(1-\lambda^2)\left(\sigma_{11} + \frac{\sigma_{22}(1+2\rho\lambda)}{(1-\lambda^2)} + \sigma_\epsilon^2\right)}. \tag{5.10}$$

Setting $\lambda = 0.75$ and $\widehat{\rho} = 0.8$, we can solve for $\rho = 0.538$. The ADF null rejection frequency for the null of I(2), i.e., the power to reject the null for the cumulated gap based on (5.2) given the specified parameterisation, setting $\sigma_{11} = \sigma_{22} = \sigma_\epsilon^2 = 1$ and undertaking $M = 10,000$ replications for $T = 100$ observations, is 83% at the 5% significance level and 42% at the 1% significance level.[2] Thus, there is a considerable probability of accepting the null of an I(2) cumulated gap, despite it being I(1).

We can derive the conditions to ensure the output gap is I(0), and thus, the cumulated gap is I(1), from (5.2):

$$y_t - \rho y_{t-1} - u_{1,t} - \epsilon_t - v_t = \lambda\left(y_{t-1} - \rho y_{t-2} - u_{1,t-1} - \epsilon_{t-1}\right), \tag{5.11}$$

such that:

$$y_t = (\rho+\lambda)\, y_{t-1} - \lambda\rho y_{t-2} + \underbrace{(u_{1,t} - \lambda u_{1,t-1}) + (\epsilon_t - \lambda\epsilon_{t-1}) + v_t}_{w_t}. \tag{5.12}$$

Reparameterising the model in differences results in:

$$\Delta y_t = (\rho + \lambda - \lambda\rho - 1)\, y_{t-1} + \lambda\rho\Delta y_{t-1} + w_t. \tag{5.13}$$

Observe that this derivation requires orthogonalisation of the error with y_{t-1} to obtain the autoregressive parameter implicit in the model. Instead, we use the estimate of ρ from (5.10) to calculate the autoregressive parameter. For an I(1) output gap, we require:

$$\rho + \lambda - \lambda\rho - 1 = 0, \tag{5.14}$$

and hence, either $\rho = 1$ or $\lambda = 1$. Under our assumption that $\lambda = 0.75$ and $\rho = 0.538$, $\rho + \lambda - \lambda\rho = 0.89$, which is close to a unit root while the individual coefficients are far

[2]Small sample critical values are obtained from Patterson (2000, Table 6.5).

from unity. Furthermore, empirical estimates of λ could be near unity given that the capital stock is I(2).

Table 5.2 records the null rejection frequencies of the ADF test under the null of an I(2) process for the cumulated y_t based on (5.2), i.e. for $\sum_{i=1}^{t} y_i$ for $t = 1, ..., T$. The ADF test includes 1 lag of the dependent variable and a restricted intercept. We set $\sigma_{11} = \sigma_{22} = \sigma_\epsilon^2 = 1$. Two sample sizes are examined, $T = 100$ and $T = 500$, with critical values given by 4.740 and 4.646 at the 5% significance level for $T = 100$ and 500, and 6.732 and 6.387 at the 1% significance level for $T = 100$ and 500, respectively. $M = 10,000$ replications are undertaken.

When either λ or ρ is equal to unity, the null hypothesis is correct. The null rejection frequency is close to the nominal size when one of the parameters is unity and the other parameter is far from unity. However, if one of ρ or λ is unity and the other is close to unity, the process is close to I(3), and hence the null rejection frequency is incorrect. For a process that is I(1) ($\rho < 1$ and $\lambda < 1$) but is close to I(2), the ability to distinguish between an I(1) and an I(2) process is very difficult in small samples. For either ρ or λ close to unity, the probability of incorrectly accepting the null hypothesis of I(2) is very high. This problem is mitigated in large samples, as the results for $T = 500$ demonstrate, but sample sizes of this magnitude are rarely available for macroeconomic data. Hence, despite an I(0) DGP, its cumulant can display characteristics close to I(2)ness in small samples, and we argue that this is plausible for the output gap such that its exclusion from the cointegrating space has a strong basis. In effect, the systematic measurement errors in the gap are already a serious difficulty, well-known to policy-makers like the MPC, and adding these cumulated to the model is likely to swamp the variability of other I(1) effects.[3]

5.2.3 Cointegration analysis of a partial system

For partial cointegration analysis we adopt the framework outlined in Harbo *et al.* (1998). Consider a p-dimensional VAR with linear deterministic terms for $\mathbf{x}_t$, where $\mathbf{x}_t$ is a $(p \times 1)$

[3]These conclusions were confirmed by inclusion of the cumulated gap in the cointegrating space as the subsequent results proved uninterpretable, delivering two cointegrating vectors in which one was I(1), based on ADF tests.

Table 5.2: Null rejection frequencies of the Augmented Dickey-Fuller test for H_0 : I(2)

	5%					*1%*				
$\lambda\backslash\rho$	*0.6*	*0.7*	*0.8*	*0.9*	*1.0*	*0.6*	*0.7*	*0.8*	*0.9*	*1.0*
$T = 100$										
0.6	0.901	0.693	0.358	0.089	0.053	0.585	0.308	0.100	0.015	0.013
0.7	0.737	0.492	0.218	0.052	0.067	0.353	0.160	0.044	0.007	0.021
0.8	0.429	0.237	0.088	0.023	0.103	0.131	0.051	0.014	0.003	0.041
0.9	0.114	0.057	0.024	0.011	0.197	0.024	0.008	0.003	0.001	0.108
1.0	0.051	0.066	0.109	0.197	-	0.012	0.019	0.042	0.108	-
$T = 500$										
0.6	1.000	1.000	1.000	0.999	0.052	1.000	1.000	1.000	0.951	0.014
0.7	1.000	1.000	1.000	0.981	0.065	1.000	1.000	1.000	0.800	0.022
0.8	1.000	1.000	1.000	0.800	0.097	1.000	1.000	0.990	0.355	0.040
0.9	1.000	0.988	0.796	0.150	0.172	0.984	0.834	0.350	0.016	0.096
1.0	0.052	0.066	0.105	0.184	-	0.013	0.024	0.045	0.104	-

Notes: ADF test based on a restricted constant and one lag of the dependent variable.

vector of variables at time t:

$$\Delta \mathbf{x}_t = \mathbf{\Pi} \mathbf{x}_{t-1} + \sum_{i=1}^{k-1} \mathbf{\Gamma}_i \Delta \mathbf{x}_{t-i} + \boldsymbol{\mu} + \boldsymbol{\delta} t + \boldsymbol{\epsilon}_t, \quad \boldsymbol{\epsilon}_t \sim \mathsf{IN}_p \left[\mathbf{0}, \mathbf{\Sigma} \right], \tag{5.15}$$

for $t = 1, ..., T$. The starting values, $(\mathbf{x}_{1-k}, ..., \mathbf{x}_0)$ are fixed, $\mathbf{\Gamma}_i$ are $(p \times p)$ matrices and $\mathbf{\Pi} = \boldsymbol{\alpha}\boldsymbol{\beta}'$, where $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are $(p \times r)$ matrices of full rank. For I(1) cointegration analysis we require the roots of the characteristic polynomial to lie on or outside the unit circle:

$$A(z) = (1 - z)\mathbf{I}_p - \mathbf{\Pi} z - \sum_{i=1}^{k-1} \Gamma_i (1 - z) z^i.$$

We also require the reduced rank condition for $\mathbf{x}_t$ to be I(1) with r cointegrating vectors given by:

$$\text{rank}\left(\boldsymbol{\alpha}'_\perp \mathbf{\Gamma} \boldsymbol{\beta}_\perp\right) = p - r,$$

where $\boldsymbol{\alpha}_\perp$ and $\boldsymbol{\beta}_\perp$ are orthogonal complements defined as $[p \times (p - r)]$ matrices such that $\boldsymbol{\alpha}'\boldsymbol{\alpha}_\perp = \mathbf{0}$ and $(\boldsymbol{\alpha}, \boldsymbol{\alpha}_\perp)$ has full rank, and similarly $(\boldsymbol{\alpha}'_\perp \boldsymbol{\beta}_\perp)$ has full rank.

To assess the partial model, we decompose $\mathbf{x}_t$ into $\mathbf{y}_t$ of dimension p_1 and $\mathbf{z}_t$ of dimension p_2: $\mathbf{x}'_t = (\mathbf{y}'_t, \mathbf{z}'_t)$. The parameter and error terms are decomposed similarly:

$$\boldsymbol{\alpha} = \begin{pmatrix} \boldsymbol{\alpha}_y \\ \boldsymbol{\alpha}_z \end{pmatrix}, \ \mathbf{\Gamma}_i = \begin{pmatrix} \mathbf{\Gamma}_{y,i} \\ \mathbf{\Gamma}_{z,i} \end{pmatrix}, \ \boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_y \\ \boldsymbol{\mu}_z \end{pmatrix}, \ \boldsymbol{\delta} = \begin{pmatrix} \boldsymbol{\delta}_y \\ \boldsymbol{\delta}_z \end{pmatrix}, \ \boldsymbol{\epsilon}_t = \begin{pmatrix} \boldsymbol{\epsilon}_{y,t} \\ \boldsymbol{\epsilon}_{z,t} \end{pmatrix}, \ \mathbf{\Sigma} = \begin{pmatrix} \mathbf{\Sigma}_{yy} & \mathbf{\Sigma}_{yz} \\ \mathbf{\Sigma}_{zy} & \mathbf{\Sigma}_{zz} \end{pmatrix}.$$

The conditional model for Δy_t is given by the equation:

$$\begin{aligned}\Delta y_t &= \boldsymbol{\omega}\Delta\mathbf{z}_t + (\boldsymbol{\alpha}_y - \boldsymbol{\omega}\boldsymbol{\alpha}_z)\,\boldsymbol{\beta}'\mathbf{x}_{t-1} + \sum_{i=1}^{k-1}(\boldsymbol{\Gamma}_{y,i} - \boldsymbol{\omega}\boldsymbol{\Gamma}_{z,i})\,\Delta\mathbf{x}_{t-i} + \left(\boldsymbol{\mu}_y - \boldsymbol{\omega}\boldsymbol{\mu}_z\right) \\ &\quad + (\boldsymbol{\delta}_y - \boldsymbol{\omega}\boldsymbol{\delta}_z)\,t + (\boldsymbol{\epsilon}_{y,t} - \boldsymbol{\omega}\boldsymbol{\epsilon}_{z,t}),\end{aligned} \tag{5.16}$$

where $\boldsymbol{\omega} = \boldsymbol{\Sigma}_{yz}\boldsymbol{\Sigma}_{zz}^{-1}$. The marginal model for $\mathbf{z}_t$ is given by:

$$\Delta\mathbf{z}_t = \boldsymbol{\alpha}_z\boldsymbol{\beta}'\mathbf{x}_{t-1} + \sum_{i=1}^{k-1}\boldsymbol{\Gamma}_{z,i}\Delta\mathbf{x}_{t-i} + \boldsymbol{\mu}_z + \boldsymbol{\delta}_z t + \boldsymbol{\epsilon}_{z,t}. \tag{5.17}$$

If $\boldsymbol{\alpha}_z = \mathbf{0}$, the long-run parameters in the conditional and marginal models are variation free. In this case, $\Delta\mathbf{z}_t$ is said to be weakly exogenous and $\boldsymbol{\beta}$ can be estimated efficiently from the conditional model.[4] Observe that other failures of weak exogeneity can arise; the formal definition of weak exogeneity given by Engle *et al.* (1983) is that the parameters of interest are a function of the conditional distribution alone, and that the parameters of the conditional and marginal distributions are variation free.

We include a constant, trend and indicator variables in the analysis. Hence, the model is specified by (5.18), in which the contemporaneous output gap is excluded for forecasting purposes. The level of the exogenous variable is included as it is assumed to be an I(0) variable. The trend is restricted to lie in the cointegrating space allowing a linear trend in the cointegrating relations, whereas the constant and dummies are unrestricted. The impact of the dummies will be negligible as long as $\left(1/\sqrt{T}\right)\boldsymbol{\phi}\sum_{i=1}^{t} D_i \to 0$ for $T \to \infty$. Subscripts correspond to the parameters in (5.16).

$$\Delta\mathbf{y}_t = (\boldsymbol{\Pi}_y, \boldsymbol{\Pi}_{y,l})\begin{pmatrix}\mathbf{y}_{t-1}\\ t\end{pmatrix} + \sum_{i=1}^{k-1}\boldsymbol{\Gamma}_{yi}\Delta\mathbf{y}_{t-i} + \sum_{i=1}^{l}\boldsymbol{\Gamma}_{zi}z_{t-i} + \boldsymbol{\mu}_x + \boldsymbol{\phi}D_t + \boldsymbol{\epsilon}_{xt}. \tag{5.18}$$

[4] Rahbek and Mosconi (1999) suggest that, rather than using the general case in which exogenous stationary regressors are added to the model given by:

$$H(r): \Delta\mathbf{y}_t = \boldsymbol{\alpha}\boldsymbol{\beta}'\mathbf{y}_{t-1} + \sum_{i=1}^{k-1}\boldsymbol{\Gamma}_i\Delta\mathbf{y}_{t-i} + \sum_{j=0}^{l}\boldsymbol{\Psi}_j\mathbf{z}_{t-j} + \boldsymbol{\mu} + \boldsymbol{\delta}t + \boldsymbol{\epsilon}_t,$$

we can extend the model to give:

$$\mathsf{H}^*(r): \Delta y_t = \boldsymbol{\alpha}\left(\boldsymbol{\beta}'y_{t-1} + \boldsymbol{\beta}_z'\sum_{i=1}^{t}\mathbf{z}_i + \boldsymbol{\beta}_l' t\right) + \boldsymbol{\Phi}\mathbf{z}_t + \sum_{i=1}^{k-1}\boldsymbol{\Gamma}_i\Delta\mathbf{y}_{t-i} + \sum_{j=0}^{l-1}\boldsymbol{\Phi}_i\Delta\mathbf{z}_{t-i} + \boldsymbol{\mu} + \boldsymbol{\epsilon}_t,$$

where $\boldsymbol{\Phi} = \sum_{j=0}^{l}\boldsymbol{\Psi}_j$ and $\boldsymbol{\Phi}_i = -\sum_{j=i+1}^{l}\boldsymbol{\Psi}_j$, $i = 0, ..., l-1$. This results in a more general model as the cumulated $\mathbf{z}_t$ appears in both the common trends and cointegrating relations. We do not adopt this approach because of the measurement errors in the cumulated output gap, but it should be noted that the critical values correctly apply to this model.

Table 5.3: Descriptive statistics for quarterly inflation

	unconditional	*-dummies*	*1966q1-1992q4*	*1993q1-1998q2*
Mean	0.0185	0.0186	0.0209	0.0068
St. dev.	0.0154	0.0012	0.0157	0.0050

5.2.4 Empirical analysis

In the empirical analysis the initial unrestricted VAR for $T = 1966q1 - 1998q2$ is defined as:

$$\mathbf{y}_t = [p_t, c_t^*, ppi_t, pw_t, e_t]' .$$

where the data are outlined in Chapter 4, (see Appendix 4.A for details). The VAR is augmented by y_{t-1}^d, and five indicator variables are included to account for special events. These are given by:

$$D_t = [I_{74:1}, I_{79:2}, D_{73:1,75:1}, D_{73:2,79:3}, D_{83:1,92:4}]' .$$

$I_{74:1}$ is an impulse dummy controlling for fluctuations in world prices, producer prices and the price level arising at the time of the first oil shock. $I_{79:2}$ is another impulse dummy accounting for the rise in VAT. $D_{73:1,75:1}$ is a blip dummy taking the values -1,1, to account for the political effects of the Heath fiscal expansion and the Wilson-Callaghan 'social contract' applied to labour market bargaining. $D_{73:2,79:3}$ is a blip dummy taking the values -1,1, accounting the measurement error and the Thatcher VAT increase and finally $D_{83:1,92:4}$ takes the value 1 to control for exchange rate fluctuations following a sharp decline in the exchange rate in 1983q1 and the fall out of the ERM in 1992. Table 5.3 provides descriptive statistics of quarterly inflation to show the shift towards lower inflation during the inflation targeting regime.

The lag length for the unrestricted VAR is determined by sequential tests of model reduction, commencing from the longest lag length of four. Table 5.4 records the F-tests of model reduction. Two statistics are reported, including the F-test against the previous lag and the F-test against the longest lag, i.e. $p = 4$. Information criteria are also reported. Eliminating the fourth lag is accepted, but further lag reductions are rejected. A lag specification of three is supported by the Akaike information criterion, although both the Schwarz and Hannan-Quinn criteria select more parsimonious models.

Table 5.4: Determination of the VAR lag length

	F (p to $p-1$)	F (4 to $p-1$)	*SIC*	*HQ*	*AIC*
$p=4$	-	-	-27.564	-29.679	-31.123
$p=3$	1.465 [0.074]	1.465 [0.074]	-28.186	-29.938	-31.132
$p=2$	1.705 [0.021]*	1.603 [0.008]**	-28.771	-30.157	-31.103
$p=1$	3.797 [0.000]**	2.394 [0.000]**	-28.886	-29.908	-30.605
$p=0$	744.5 [0.000]**	114.9 [0.000]**	-15.053	-15.709	-16.158

Notes: F-test of model reduction, where p = lag length of VAR, with p-values reported in brackets. * and ** indicate significance at the 5% and 1% significance level respectively.

Table 5.5: Tests for mis-specification of the unrestricted VAR(3)

	AR($1-4$) F(4,102), F(100,404)	ARCH(4) F(4,98)	Normality $\chi^2(2)$, $\chi^2(10)$	Hetero F(32,73), F(480,863)
p_t	1.801 [0.134]	0.937 [0.446]	1.047 [0.593]	0.981 [0.510]
c_t^*	1.890 [0.118]	1.194 [0.319]	1.591 [0.451]	0.947 [0.557]
ppi_t	0.908 [0.462]	0.612 [0.655]	0.768 [0.681]	1.383 [0.128]
pw_t	2.653* [0.037]	0.339 [0.851]	0.360 [0.835]	1.145 [0.312]
e_t	0.872 [0.483]	1.280 [0.283]	3.269 [0.195]	0.786 [0.772]
Multivariate tests	1.283* [0.050]	...	10.141 [0.428]	0.746 [0.999]

Notes: Tests include up to fourth order autocorrelation, fourth order ARCH, normality and heteroskedasticity. Both single and multivariate tests are reported, with p−values given in brackets. * denotes significance at the 5% level.

Tests on lags of the exogenous variable y^d indicate that only one lag is required.

Tests for evidence of mis-specification for the unrestricted VAR with a lag length of three are reported in Table 5.5, with a residual analysis of the VAR recorded in Figure 5.2. The tests are, in general, satisfactory, although there is some evidence of autocorrelation in world prices. This implies that the multivariate test fails residual correlation up to the fourth order at the 5% significance level. Including additional lags does not solve the problem.[5] The residuals and their corresponding densities, correlograms and QQ-plots support the evidence for a reasonably well-specified model. The recursive graphics are also examined to ensure that there is no evidence of parameter change. Overall, the unrestricted VAR is congruent and we proceed with the analysis.

[5]Scheffe's S procedure (see Savin, 1984) can be used for multiple testing, commencing with the overall F-statistic for the multivariate test and proceeding to the individual tests, allowing for all combinatorial comparisons. Scheffe's procedure tends to be conservative, and multiple testing procedures can deliver contrasting results, implying that the interpretation of the multivariate and single-equation test results is not straightforward if disparities emerge.

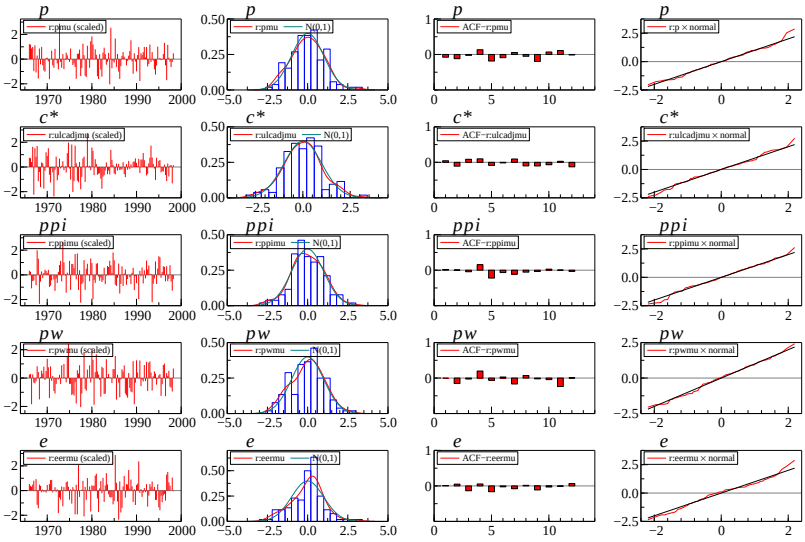

Figure 5.2: Residual analysis of the unrestricted VAR, including scaled residuals, residual densities and histograms, residual correlograms, and QQ plots.

Table 5.6: I(1) cointegration analysis

r	*0*	*1*	*2*	*3*	*4*	*5*
l	2097.81	2131.37	2146.401	2154.95	2160.20	2161.30
λ		0.403	0.207	0.123	0.078	0.017
	$r = 0$	$r \leq 1$	$r \leq 2$	$r \leq 3$	$r \leq 4$	
$Q(r)$	126.98**	59.85	29.78	12.70	2.19	
p−val	0.000**	0.103	0.521	0.762	0.942	

Notes: Log-likelihoods (l), eigenvalues (λ), trace statistics ($Q(r)$) and p−values of the trace statistic (p−val) for all possible cointegrating ranks.

5.2.5 I(1) analysis

To determine the cointegrating rank, Table 5.6 reports the log-likelihood values (l), eigenvalues (λ), and trace statistics ($Q(r)$) for the VAR(3). Standard p−values are reported, based on Doornik (1998), although we focus on the critical values for the partial cointegration model given by Harbo *et al.* (1998), which are reported in Table 5.7. The roots of the companion matrix indicate that there is no eigenvalue > 1, which suggests that there is no explosive root. We do not consider an I(2) analysis as inference using standard critical values cannot be made due to the partial model in which there are nuisance parameters. We work within the I(1) framework and conclude that there is one cointegrating vector.

To identify the long-run structure within the model, the unrestricted factor loadings and cointegrating relations are reported in (5.19), in which we normalise on p.

Table 5.7: Cointegration analysis critical values for a partial system

Critical Values				
50%	80%	90%	95%	97.5%
54.8	63.0	67.9	71.7	75.2

Notes: Critical values from Harbo *et al.* (1998). Quantiles for $r = 1$ with one exogenous variable and five endogenous variables.

Table 5.8: Tests for weak exogeneity

	Test statistic	p−*value*
p_t	24.616	0.000^{**}
c_t^*	0.041	0.839
ppi_t	8.481	0.004^{**}
e_t	14.089	0.002^{**}
pw_t	0.578	0.447

Notes: The test statistic is distributed as $\chi^2(1)$ under the assumption that the correct cointegrating rank is imposed.

$$\boldsymbol{\alpha\beta'} = \begin{pmatrix} -0.282 \\ (0.049) \\ -0.018 \\ (0.198) \\ -0.670 \\ (0.225) \\ 0.849 \\ (0.025) \\ -0.020 \\ (0.025) \end{pmatrix} \begin{pmatrix} 1 & -0.873 & -0.094 & -0.050 & 0.039 & -0.0012 \\ \cdots & (0.058) & (0.025) & (0.030) & (0.104) & (0.0006) \end{pmatrix}. \tag{5.19}$$

We test the restriction that the coefficient on world prices is equal to the negative coefficient on the exchange rate. The restriction is accepted; $\chi^2(1) = 0.0107[0.918]$; giving world prices in sterling.[6] We also test the restriction that the trend does not enter the cointegrating vector. This restriction is marginally accepted; $\chi^2(1) = 3.6253[0.057]$; and we impose a zero coefficient on the trend. Furthermore, tests of weak exogeneity, which imply a zero coefficient on the α parameter, are reported in Table 5.8. We accept the weak exogeneity hypothesis for unit labour costs and world prices. Imposing all restrictions is accepted; $\chi^2(4) = 5.5150[0.238]$; and the resulting restricted estimate of $\alpha\beta'$ is given by:

$$\boldsymbol{\alpha\beta'} = \begin{pmatrix} -0.228 \\ (0.043) \\ 0 \\ -0.711 \\ (0.170) \\ 0.789 \\ (0.216) \\ 0 \end{pmatrix} \begin{pmatrix} 1 & -0.925 & -0.048 & -0.024 & 0.024 \\ \cdots & (0.022) & (0.015) & (0.012) & \cdots \end{pmatrix}. \tag{5.20}$$

[6] [.] reports p−values.

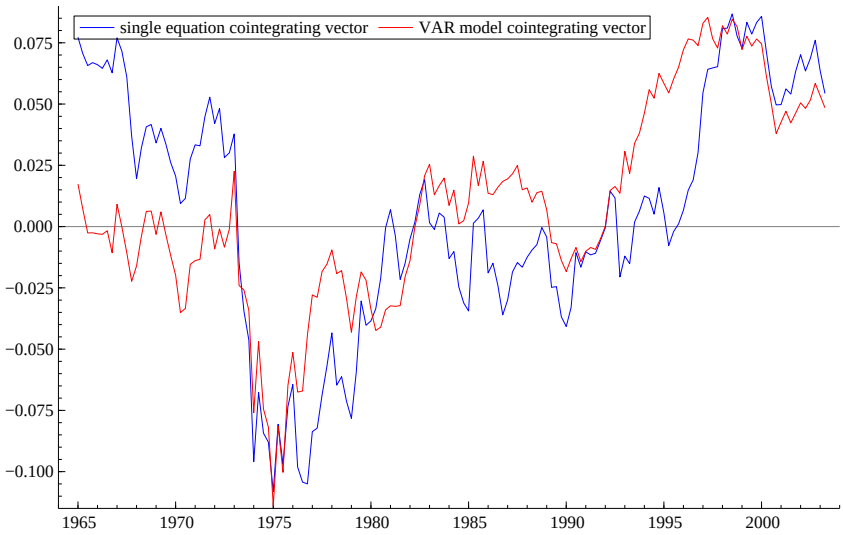

Figure 5.3: The VAR cointegrating vector and the single-equation counterpart.

This is analogous to the single-equation mark-up found in (4.19). The trend is insignificant in both the VAR cointegrating relation and the single-equation model. The coefficient on unit labour costs is larger than in the single-equation model and the coefficient on world prices in sterling is much smaller. The cointegrating vector is recorded in Figure 5.3, along with the single-equation mark-up in (4.19). Both cointegrating relations have a zero mean for the in-sample period.

We next map the data to I(0) space to forecast inflation. In order to implement the reductions to a more parsimonious model, we use the reduced form VAR framework in PcGets. This algorithm implements the *Gets* selection strategy analogous to that used in the single-equation model but applied to a multivariate GUM, see Hendry and Krolzig (2001, ch.8) for a discussion, although currently the algorithm is applied individually to each equation and so omits cross-equation restrictions. The GUM is in VEqCM form, containing 18 regressors in each equation, including two lags of the differenced variables, the lagged cointegrating vector, an intercept, and five dummies. The dominant congruent model is derived, reported in equations (5.21) to (5.24). The test of model reduction from the full VEqCM to the model specification outlined is accepted; $\chi^2(55) = 68.358[0.107]$. Figure 5.4 records the fitted and actual values, residuals, density, and correlogram for quarterly inflation; the variable of focus in the forecasting exercise.

There is evidence of autocorrelation within the system that is not removed by including extra lags, hence we report HACSE estimates (see Andrews, 1991). If the equations are estimated independently there is no evidence of autocorrelation. The model represents a reasonable fit. The equation standard error of 0.72% is larger than the single-equation model with a standard error of 0.61%, due to the smaller variable set and restricted dynamics. The Chow test is accepted; $\mathsf{F}_{\mathsf{Chow}}(100, 123) = 0.851\,[0.798]$.

$$\begin{aligned}
\Delta p_t &= \underset{(0.156)}{0.496}\Delta pw_{t-1} + \underset{(0.161)}{0.647}\Delta pw_{t-2} + \underset{(0.035)}{0.168}y^d_{t-1} + \underset{(0.044)}{0.157}\Delta c^*_{t-1} \\
&\quad \underset{(0.017)}{-0.045}cv_{t-1} - \underset{(0.007)}{0.028}I_{74:1} + \underset{(0.005)}{0.043}D_{73:2,79:3} + \underset{(0.005)}{0.014}D_{73:1,75:1} \qquad (5.21)\\
\widehat{\sigma} &= 0.718\% \quad \mathsf{F}_{\mathsf{ar}}(5, 107) = 5.780^{**} \quad \mathsf{F}_{\mathsf{het}}(31, 91) = 0.778 \\
\chi^2_{\mathsf{nd}}(2) &= 0.685 \quad \mathsf{F}_{\mathsf{arch}}(4, 115) = 0.482 \quad T = 1966q1 - 1998q2.
\end{aligned}$$

$$\begin{aligned}
\Delta c^*_t &= \underset{(0.083)}{0.231}\Delta p_{t-1} + \underset{(0.053)}{0.338}y^d_{t-1} + \underset{(0.062)}{0.120}\Delta c^*_{t-2} + \underset{(0.145)}{1.028}\Delta pw_{t-1} - \underset{(0.028)}{0.108}\Delta ppi_{t-1} \\
&\quad \underset{(0.028)}{+0.065}\Delta ppi_{t-2} + \underset{(0.007)}{0.050}D_{73:1,75:1} + \underset{(0.007)}{0.019}D_{73:2,79:3} \qquad (5.22)\\
\widehat{\sigma} &= 1.030\% \quad \mathsf{F}_{\mathsf{ar}}(5, 107) = 2.372^{*} \quad \mathsf{F}_{\mathsf{het}}(31, 91) = 1.452 \\
\chi^2_{\mathsf{nd}}(2) &= 0.947 \quad \mathsf{F}_{\mathsf{arch}}(4, 115) = 1.365 \quad T = 1966q1 - 1998q2.
\end{aligned}$$

$$\begin{aligned}
\Delta ppi_t &= \underset{(0.060)}{0.286}\Delta ppi_{t-1} + \underset{(0.116)}{0.341}y^d_{t-1} - \underset{(0.063)}{0.189}cv_{t-1} \\
&\quad \underset{(0.023)}{+0.214}I_{74:1} + \underset{(0.017)}{0.051}D_{83:1,92:4} + \underset{(0.002)}{0.008} \qquad (5.23)\\
\widehat{\sigma} &= 2.617\% \quad \mathsf{F}_{\mathsf{ar}}(5, 107) = 10.363^{**} \quad \mathsf{F}_{\mathsf{het}}(31, 91) = 4.760^{**} \\
\chi^2_{\mathsf{nd}}(2) &= 5.568 \quad \mathsf{F}_{\mathsf{arch}}(4, 115) = 5.281^{**} \quad T = 1966q1 - 1998q2.
\end{aligned}$$

$$\begin{aligned}
\Delta pw_t &= \underset{(0.022)}{0.076}\Delta p_{t-2} + \underset{(0.016)}{0.040}y^d_{t-1} + \underset{(0.077)}{0.477}\Delta pw_{t-1} + \underset{(0.070)}{0.253}\Delta pw_{t-2} \\
&\quad \underset{(0.009)}{+0.029}\Delta ppi_{t-1} + \underset{(0.003)}{0.009}I_{74:1} + \underset{(0.003)}{0.010}I_{79:2} + \underset{(0.0005)}{0.002} \qquad (5.24)\\
\widehat{\sigma} &= 0.315\% \quad \mathsf{F}_{\mathsf{ar}}(5, 107) = 4.492^{**} \quad \mathsf{F}_{\mathsf{het}}(31, 91) = 1.444 \\
\chi^2_{\mathsf{nd}}(2) &= 0.754 \quad \mathsf{F}_{\mathsf{arch}}(4, 115) = 0.546 \quad T = 1966q1 - 1998q2.
\end{aligned}$$

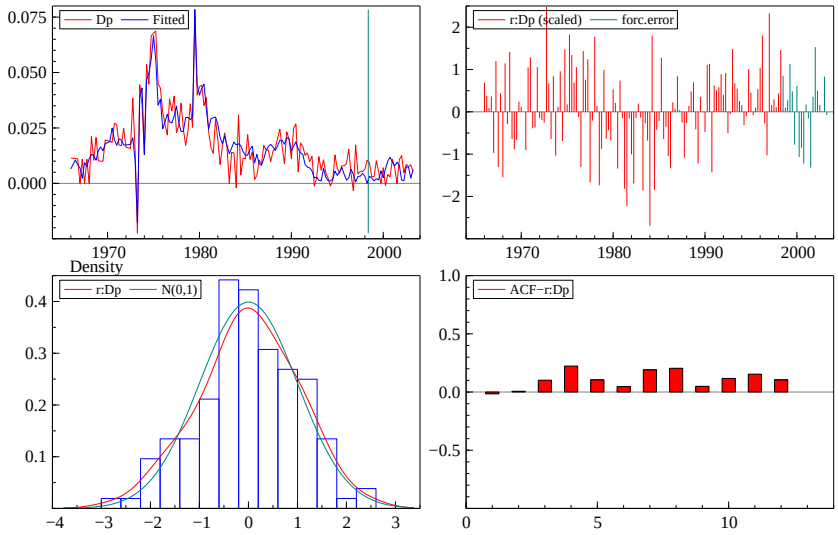

Figure 5.4: Fitted and actual values, residuals, density, and correlogram for the VEqCM model of quarterly inflation.

$$\begin{aligned}
\Delta e_t &= \underset{(0.412)}{1.219}\Delta pw_{t-2} + \underset{(0.094)}{0.400}cv_{t-1} + \underset{(0.024)}{0.075}I_{79:2} - \underset{(0.020)}{0.113}D_{83:1,92:4} - \underset{(0.006)}{0.021} \qquad (5.25)\\
\widehat{\sigma} &= 2.889\% \quad \mathsf{F}_{\mathsf{ar}}(5,107) = 5.265^{**} \quad \mathsf{F}_{\mathsf{het}}(31,91) = 0.595\\
\chi^2_{\mathsf{nd}}(2) &= 3.270 \quad \mathsf{F}_{\mathsf{arch}}(4,115) = 1.238 \quad T = 1966q1 - 1998q2.
\end{aligned}$$

There is a substantial impact from external prices feeding through to domestic inflation with both lags corresponding to a coefficient of approximately unity. This may well be reflecting collinearity with domestic prices as well, as the lagged dependent variable is not retained. The output gap has a 17% effect, so is an important source of inflation. The cointegrating vector enters with an impact of 5%, and so the exchange rate, unit labour costs and producer prices all enter significantly. This is much smaller than the 13% impact found in the single-equation model, and is indeed much smaller than (5.20) would suggest. An intercept is not retained, suggesting that there is no evidence of autonomous inflation. The output gap feeds into unit labour costs and producer prices with a greater impact than inflation at 34%, suggesting that excess demand pressures are recognised more fully at the intermediate stage. The cointegrating vector does not enter into the equation for unit labour costs, reflecting the evidence found for weak exogeneity. External inflation has a substantial impact on unit labour costs, with a coefficient of approximately one. The exchange rate equation contains the lagged external inflation rate

and the mark-up which is difficult to explain, but an equation standard error of 2.89% suggests that the model is poor. The exchange rate is difficult to model; often a random walk provides the best fit, although it is marginally inferior to (5.25) for this sample.

5.3 Quarterly inflation forecasts

To assess the forecast performance of the inflation models, we undertake a forecast comparison exercise in which their *ex ante* forecasts are compared with the forecasts from alternative forecasting devices. Conventional forecasting theory from the seminal works of Box and Jenkins (1970), Klein (1971), and Granger and Newbold (1986) assume that the best in-sample model is the best out-of-sample model, delivering the lowest MSFE matrix. Hence, we would assume the econometric models developed would produce the best forecasts. However, the empirical forecasts produced in this section show that there is no definitive ranking of forecasting models, and often models with little or no economic content can outperform well-specified econometric models. Section 5.3.1 outlines the forecasting methods and evaluation criteria, section 5.3.2 discusses the forecasting models and section 5.3.3 reports the results for quarterly inflation.

5.3.1 Forecasting methods

Estimating a model over $t = 1, ..., T$, with a forecast horizon of $t = T + 1, ..., T + H$, the forecast at $T + h$ is given in (5.26), where $\mathcal{I}_T$ is the information set at time T, $\boldsymbol{\theta}_T$ is the set of estimated model parameters at time T, and the forecast is a function h steps ahead, ψ_h. The resulting forecast error in period $T + h$ is given in (5.27).

$$\widehat{y}_{T+h|T} = \psi_h \left(\mathcal{I}_T, \widehat{\boldsymbol{\theta}}_T \right) \tag{5.26}$$

$$e_{T+h|t} = y_{T+h|T} - \widehat{y}_{T+h|T}. \tag{5.27}$$

To compare the accuracy of forecasts we shall examine the bias (ME), efficiency (MAE), and a composite of these (RMSFE), for the forecasts derived from each model.[7] As forecast accuracy rankings can change over different forecast horizons (based on

[7]The MAPE is not reported for quarterly inflation as MAPE $= \infty$ for $y_t = 0$, and two of the forecast horizon realisations of $\Delta p_{T+h} = 0$.

MSFE), multi-step forecasts are also examined for 4-steps ahead. Clements and Hendry (1993a, 1993b) demonstrate that MSFEs are not invariant under non-singular, scale-preserving linear transformations, and instead propose the GFESM criterion (see Chapter 1), although as we compare many forecast models the calculation of all intermediate forecast horizons is computationally intensive, and consequently we report RMSFE instead of GFESM, but the caveats of such an evaluation criterion must be noted.

Multi-step forecasting techniques, including the iterated 1-step estimator and the direct h-step estimator, are outlined in (1.17) and (1.18). Only the direct h-step forecast can be used in the single-equation framework, and so a model must be developed to produce direct 4-step ahead forecasts. The *Gets* technique used in Chapter 4 is applied to derive a model of quarterly inflation using information lagged one year. The initial GUM contains lags 4 to 7 of y^d, U^d, e_r, $(c^* - p)$, s, $(ppi - p)$, $(rent - p)$, $(imp - p)$, $(oil - p)$, $(n - p)$, R_l, Δppi, $\Delta rent$, Δimp, Δoil, Δp, Δc^*, $\Delta m4$, Δn, ΔR_s, ΔR_l, Δpw, lags 4 and 5 of $\Delta e_t e_{r,t-j}$ and $\Delta e_t e^2_{r,t-j}$, and an intercept, trend, and indicators.[8] As information in periods $t-1$ to $t-3$ are excluded, residual autocorrelation is likely to bias the estimated coefficient standard errors. In order to overcome this, looser significance levels are used to select the model (using the expert-users strategy in PcGets) and the autocorrelation mis-specification test is excluded from the test battery. The resulting specific model is estimated in PcGive to establish HACSEs based on Andrews (1991), reported in parentheses.[9]

$$\begin{aligned}
\Delta p_t &= \underset{(0.144)}{0.329}\Delta pw_{t-4} + \underset{(0.015)}{0.086}\Delta ppi_{t-4} + \underset{(0.030)}{0.298}y^d_{t-4} - \underset{(0.028)}{0.120}U^d_{t-7} \\
&\quad + \underset{(0.004)}{0.016}\Delta oil_{t-5} + \underset{(0.052)}{0.157}\Delta rent_{t-4} - \underset{(0.022)}{0.060}\pi^*_{t-4} + \underset{(0.002)}{0.051}D_{73:2,79:3} \\
&\quad + \underset{(0.003)}{0.025}D_{72:4,74:1} + \underset{(0.002)}{0.021}D_{84:1,84:2} + \underset{(0.0018)}{0.0075} \qquad (5.28) \\
\mathsf{R}^2 &= 0.814 \quad \widehat{\sigma} = 0.698\% \quad SIC = -9.597 \\
\mathsf{F}_{\mathsf{ar}}(5,110) &= 0.884 \ \mathsf{F}_{\mathsf{arch}}(4,107) = 0.990 \quad \mathsf{F}_{\mathsf{het}}(20,94) = 0.642 \quad \chi^2_{\mathsf{nd}}(2) = 2.239 \\
\mathsf{F}_{\mathsf{reset}}(1,114) &= 2.703 \quad \mathsf{F}_{\mathsf{Chow}}(20,115) = 0.991 \quad T = 1967q1 - 1998q2.
\end{aligned}$$

[8] The indicators included are those reported in Table 4.1.

[9] In practice, further selection may be undertaken if the t-statistics from the HACSE estimates result in insignificant coefficients, and we use an iterative procedure between PcGive and PcGets to determine the appropriate significance levels for selection, as opposed to resorting to stepwise regression.

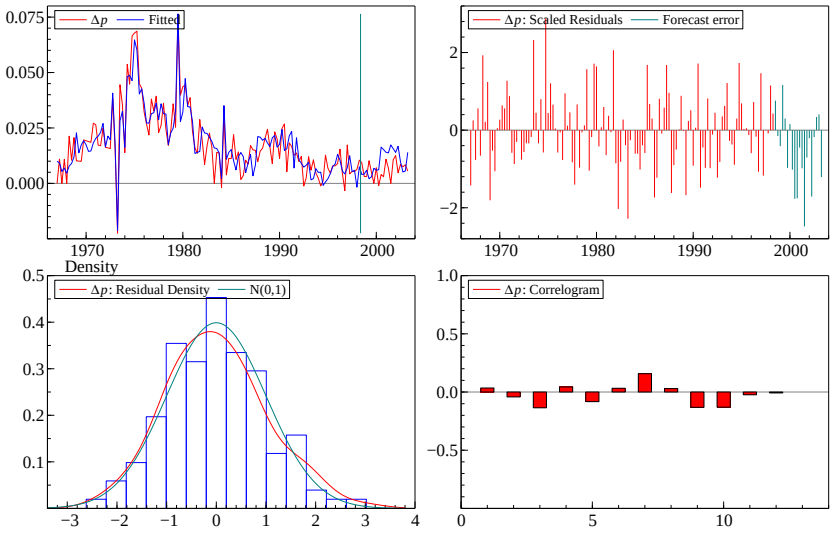

Figure 5.5: Fitted and actual values, residuals, density, and correlogram for the EqCM model of quarterly inflation to forecast 4-steps ahead.

The model is similar to the quarterly model, with a remarkably small rise in the equation standard error from 0.61% to 0.70% due to the reduced information set. Neither the acceleration of inflation nor the PPP interaction term enter the model, implying these effects are short-term occurring within one year. The intercept is larger at 0.0075, and the impact of the mark-up is much smaller as it is lagged by four periods. Even with the reduced information set, the model passes all diagnostics, including autocorrelation and heteroskedasticity, and the marginal decline in the equation standard error suggests that inflation may well correspond to an annual process because of time lags, where changes in input prices feed through to final prices slowly due to menu costs etc. The model fit is recorded in Figure 5.5, along with the scaled residuals, their correlogram and residual density.

5.3.2 Forecasting models

A range of forecasting models is assessed, including the single-equation and vector equilibrium correction models in equations (4.20) and (5.21). These econometric models are compared with various robust forecasting devices including a differenced VAR based on a five-year rolling average growth rate (DV); a differenced EqCM (ΔEqCM) and VEqCM (ΔVEqCM), in which the coefficients from the equilibrium correction models are

imposed; differenced EqCM and VEqCM equations that exclude the I(-1) double differenced terms and therefore just consist of the differenced cointegrating vector (ΔEqCM$_\beta$, ΔVEqCM$_\beta$); a random walk, or equivalently a double differenced VAR (DDV); a rapid updating device, given by a smoothed difference (SMD); an autoregressive model (AR); and a pooled forecast (Pool).

Clements and Hendry (1998b, 1999) demonstrate that non-causal models can outperform well-specified econometric models in a forecasting context if there are non-modelled structural breaks, as non-causal models may be more robust to breaks in the deterministic terms. We shall briefly outline the forecasting models using a simplified VAR(1) for exposition, numbering the models for clarity.

1. Assume $\mathbf{x}_t = \boldsymbol{\tau} + \Gamma \mathbf{x}_{t-1} + \boldsymbol{\epsilon}_t$. Taking differences results in:

$$\begin{aligned} \Delta \mathbf{x}_t &= \boldsymbol{\alpha}\boldsymbol{\beta}' \mathbf{x}_{t-1} + \boldsymbol{\tau} + \boldsymbol{\epsilon}_t & (5.29) \\ &= \boldsymbol{\gamma} + \boldsymbol{\alpha}\left(\boldsymbol{\beta}' \mathbf{x}_{t-1} - \boldsymbol{\mu}\right) + \boldsymbol{\epsilon}_t. & (5.30) \end{aligned}$$

   where $\mathbf{\Gamma} = \mathbf{I} + \boldsymbol{\alpha}\boldsymbol{\beta}'$ and $\boldsymbol{\tau} = \boldsymbol{\gamma} - \boldsymbol{\alpha}\boldsymbol{\mu}$.[10] It is clear from (5.30) that breaks occurring in the deterministic terms can arise through $\boldsymbol{\mu}$, where $\mathsf{E}\left[\boldsymbol{\beta}' \mathbf{x}_t\right] = \boldsymbol{\mu}$, or via breaks in the unconditional growth rate of the system, $\boldsymbol{\gamma}$, where $\mathsf{E}\left[\Delta \mathbf{x}_t\right] = \boldsymbol{\gamma}$.[11] Clements and Hendry (1999) outline the forecast errors and variances of (5.30) when a break occurs in $\boldsymbol{\mu}$, $\boldsymbol{\gamma}$ or $\boldsymbol{\alpha}$, along with the analogous forecast errors and variances for a variety of forecasting devices. The success of robust forecasting devices highlight the presence of breaks in the deterministic terms. Shifts in the mean are the most problematic for forecasting. If there occurs a shift in the equilibrium mean that is unaccounted for, forecasts will be adjusting to the old mean and will therefore be off-target for the entire adjustment period.

2. Using (5.30) as the in-sample DGP, the VAR in differences is given by:

$$\Delta \mathbf{x}_t = \boldsymbol{\gamma} + \boldsymbol{\xi}_t. \quad (5.31)$$

[10] To map (5.18) to this simple form we set $\mathbf{\Pi}_{y,l} = \mathbf{0}$, $\mathbf{\Gamma}_{yi} = \mathbf{0}$ for $i = 1, ..., k-1$, $\mathbf{\Gamma}_{zi} = \mathbf{0}$ for $i = 1, ..., l$ and $\boldsymbol{\phi} = \mathbf{0}$, where $\mathbf{\Pi}_y = \boldsymbol{\alpha}\boldsymbol{\beta}'$, $\boldsymbol{\mu}_x = \boldsymbol{\tau}$ and $\boldsymbol{\epsilon}_{x,t} = \boldsymbol{\epsilon}$.

[11] A third possibility for deterministic breaks arise from shifts in $\boldsymbol{\alpha}$, the adjustment coefficients affecting $\boldsymbol{\tau}$. We abstract from this case in the following analysis.

This will be mis-specified unless the cointegrating rank is 0. The elimination of $\boldsymbol{\mu}$ and $\boldsymbol{\alpha}$ robustifies the model to changes in these parameters over the forecast horizon. In practice, a five-year rolling average is used rather than the in-sample inflation rate because of the regime changes over the sample period.[12]

3. The double differenced VAR (DDV) is given by:

$$\Delta^2 \mathbf{x}_t = \boldsymbol{\zeta}_t \tag{5.32}$$

and this will track inflation by one quarter, see (1.12). This device eliminates $\boldsymbol{\gamma}$, $\boldsymbol{\mu}$ and $\boldsymbol{\alpha}$; all the potentially damaging terms. A further advantage of differencing is that the forecast retains the structural information. We can rewrite (5.32) as:

$$\Delta \mathbf{x}_t = \Delta \mathbf{x}_{t-1} + \boldsymbol{\zeta}_t, \tag{5.33}$$

where $\Delta \mathbf{x}_{t-1} = \boldsymbol{\gamma} + \boldsymbol{\alpha}\left(\boldsymbol{\beta}' \mathbf{x}_{t-2} - \boldsymbol{\mu}\right) + \boldsymbol{\epsilon}_{t-1}$, and so, if there are no breaks, the structural model is included in the forecast (albeit with a one period lag) and the only extra cost is an additional error term, $\boldsymbol{\zeta}_t$.

4. Another adaptive device that may be used is differencing the VEqCM:

$$\Delta \mathbf{x}_t = \Delta \mathbf{x}_{t-1} + \boldsymbol{\alpha}\boldsymbol{\beta}' \Delta \mathbf{x}_{t-1} + \Delta \boldsymbol{\epsilon}_t = \left(\mathbf{I} + \boldsymbol{\alpha}\boldsymbol{\beta}'\right) \Delta \mathbf{x}_{t-1} + \boldsymbol{\nu}_t, \tag{5.34}$$

which is the first difference of the initial VAR with the rank restrictions from cointegration imposed. Alternatively, writing (5.34) as:

$$\Delta^2 \mathbf{x}_t = \boldsymbol{\alpha}\boldsymbol{\beta}' \Delta \mathbf{x}_{t-1} + \boldsymbol{\nu}_t \tag{5.35}$$

shows that the DDV can be augmented by $\boldsymbol{\alpha}\boldsymbol{\beta}' \Delta \mathbf{x}_{t-1}$. As the device differences the mean, a shift in $\boldsymbol{\mu}$ will imply the forecast will fail in the period following the break, but will then correct as $\Delta \boldsymbol{\mu} = \mathbf{0}$ in subsequent periods. Hence, as in the case of the DDV, a differenced VEqCM or EqCM will robustify forecasts to deterministic shifts. On a note of caution, unnecessary differencing will lead to increased uncertainty,

[12] Determination of the optimal in-sample period for developing a forecasting model will depend on the trade-off between the benefits of a large sample period and the increased probability of structural breaks.

which may increase the MSFE.

5. In empirical models, the VEqCM is usually augmented by differenced lagged regressors to allow for more complex dynamics, and so (5.30) would be:[13]

$$\Delta \mathbf{x}_t = \boldsymbol{\gamma} + \boldsymbol{\alpha}\left(\boldsymbol{\beta}'\mathbf{x}_{t-1} - \boldsymbol{\mu}\right) + \boldsymbol{\delta}'\Delta \mathbf{x}_{t-1} + \boldsymbol{\epsilon}_t, \tag{5.36}$$

for a VAR(2), and hence the differenced VEqCM contains I(-1) terms:

$$\Delta \mathbf{x}_t = \left(\mathbf{I} + \boldsymbol{\alpha}\boldsymbol{\beta}'\right)\Delta \mathbf{x}_{t-1} + \boldsymbol{\delta}'\Delta^2 \mathbf{x}_{t-1} + \boldsymbol{\nu}_t, \tag{5.37}$$

The ΔEqCM$_\beta$ and ΔVEqCM$_\beta$ models set $\boldsymbol{\delta} = \mathbf{0}$ to avoid inflating the MSFE by highly volatile I(-1) variables.

6. The lag length of the autoregressive model is selected using AIC. A lag length of three is selected and the model is estimated over the full in-sample period.

7. The longer period difference attempts to capture a more rapid updating of the coefficients on the deterministic terms such that structural breaks are picked up faster. Clements and Hendry (2005a) outline the ideas behind rapid updating. Assume a moving average of past actual growth rates given by $\widetilde{\boldsymbol{\theta}}_T = \frac{1}{m+1}\sum_{i=0}^{m}\Delta \mathbf{x}_{T-i}$, so that the forecast at $T+1$ is:

$$\widetilde{\Delta \mathbf{x}}_{T+1|T} = \widetilde{\boldsymbol{\theta}}_T. \tag{5.38}$$

Then:

$$(m+1)\,\widetilde{\boldsymbol{\theta}}_T = \sum_{i=0}^{m}\Delta \mathbf{x}_{T-i} = \Delta \mathbf{x}_T - \Delta \mathbf{x}_{T-(m+1)} + (m+1)\,\widetilde{\boldsymbol{\theta}}_{T-1}, \tag{5.39}$$

so:

$$\widetilde{\boldsymbol{\theta}}_T = \widetilde{\boldsymbol{\theta}}_{T-1} + \frac{1}{m+1}\Delta\Delta_{m+1}\mathbf{x}_T, \tag{5.40}$$

reflecting aspects of Kalman filtering. When m=0 we obtain the DDV, which corresponds to updating the intercept, whereas larger values of m smooth intercept estimates but adapt more slowly. Setting $m = 3$, the SMD for quarterly data is

[13] This model corresponds to $\boldsymbol{\delta} = \boldsymbol{\Gamma}_{y1}$ in (5.18).

given by:

$$\Delta \mathbf{x}_{T+i|T+i-1} = \frac{1}{4}\sum_{j=1}^{4} \Delta \mathbf{x}_{T+i-j} = \frac{1}{4}\Delta_4 \mathbf{x}_{T+i-1}. \tag{5.41}$$

8. The pooled forecast is computed as an unweighted average. Hendry and Clements (2004) show that in the presence of structural breaks, simple averaging may dominate over estimating the weights for the forecast combination.

5.3.3 Forecasting results

A comparison of the forecast results for the models considered, over the period 1998q3-2003q2, is provided in Table 5.9. This records the 1-step and 4-step ME, MAE, and RMSFE for the various forecasting models. First, assessing the 1-step forecasts, Figure 5.6a shows the reasonable performance of the VEqCM for quarterly inflation, in which all of the realised outcomes lie within the $\pm 2\widehat{\sigma}_f$ error bars.[14] The model does overpredict inflation over 2000 and 2001, with an average prediction of 3.2% p.a. compared to an actual average inflation rate of 2.1%. However, the forecast error variance is not substantial (note that all graphs are on the same scale for comparison) and the RMSFE of 0.55% is fairly good. The ΔVEqCM, imposing the coefficients from equation (5.21) are recorded in Figure 5.6b along with ΔVEqCM_β. The MAE of the ΔVEqCM compared to the ΔVEqCM_β reflects the inflated forecast variance due to the inclusion of I(-1) terms.

Comparing these results to two other robust forecasting devices, Figures 5.6c and 5.6d record the forecasts from the DDV and DV respectively. The DV is based on a 5-year rolling average mean growth rate as opposed to the sample mean (which yields a RMSFE of 1.31%). Both adaptive devices perform well, with mean errors of 0.03%. The DDV tracks inflation by one quarter, demonstrating that not only should the DDV perform well when there are breaks to adjust to, but the device is also successful when there are no breaks as the method is tracking a stable series. The DV does particularly well as inflation has been very stable over the forecast horizon, demonstrating that if breaks cannot be pre-empted and the series is stationary, a forecast based on the mean unconditional growth rate is a good model. In contrast, if inflation were I(1), as in

[14] Note that the forecast standard error bars are incorrectly estimated by PcGive when the 1-step errors are not homoskedastic innovation errors over the forecast period with constant parameters.

Table 5.9: Summary of quarterly inflation forecasts

	ME		*MAE*		*RMSFE*	
Horizon	*1-step*	*4-step*	*1-step*	*4-step*	*1-step*	*4-step*
VEqCM	0.06	-0.20	0.45	0.47	0.55	0.58
ΔVEqCM	-0.06	-0.18	0.61	0.63	0.69	0.73
ΔVEqCM$_\beta$	-0.03	-0.08	0.46	0.42	0.54	0.51
DV	-0.03	-0.03	**0.34**	**0.34**	0.41	**0.40**
DDV	-0.03	**0.01**	0.47	0.40	0.55	0.49
EqCM	-0.08	-0.38	0.42	0.60	0.54	0.75
ΔEqCM	-0.03	**-0.01**	0.68	0.65	0.79	0.73
ΔEqCM$_\beta$	-0.03	-0.02	0.46	0.38	0.55	0.47
AR(3)	-0.24	-0.53	0.39	0.55	0.47	0.66
SMD	**0.01**	0.02	**0.34**	0.35	**0.40**	0.43
Pool	-0.09	-0.12	0.37	0.37	0.43	0.43

Notes: Results for 1-step and 4-step ahead forecasts over 1998q3-2003q2. Figures reported as percentages, with the best forecasting performance highlighted in **bold**.

the 1970s and 1980s, this forecasting device would not be as successful as a further differencing using the DDV device.

Examining the performance of the single-equation model, Figure 5.7a records the forecasts from the EqCM. The RMSFE of 0.54% is comparable to that of the VEqCM. Differencing the EqCM, as shown in Figure 5.7b, does not yield any improvement as the variance is inflated, but removing the short-run dynamics does improve the forecasts. Both adaptive devices given by the AR(3) model in Figure 5.7c and the SMD in Figure 5.7d forecast particularly well. The adaptive devices result in much smoother forecasts which outperform the econometric models. The pooled forecast performs well if forecast errors offset each other. While the majority of forecasts are biased upwards over the forecast horizon, the pooled forecast gains by reducing the variance.

Figures 5.8 and 5.9 record the corresponding forecasts for one year ahead quarterly inflation. Again, the SMD and the DV predict inflation well, showing the relative stability of inflation over the forecast period; the pooled forecast is also good. There is a more marked improvement in the ΔVEqCM$_\beta$ and ΔEqCM$_\beta$ at the 4-step horizon. Unusually, almost all models forecast better out-of-sample than the VEqCM performs in-sample, with an equation standard error of 0.72% which can be directly compared to the RMSFE, again reflecting the quiescence of the economy over the forecast horizon.

Table 5.10 examines the dynamic inflation forecasts over the entire forecast horizon

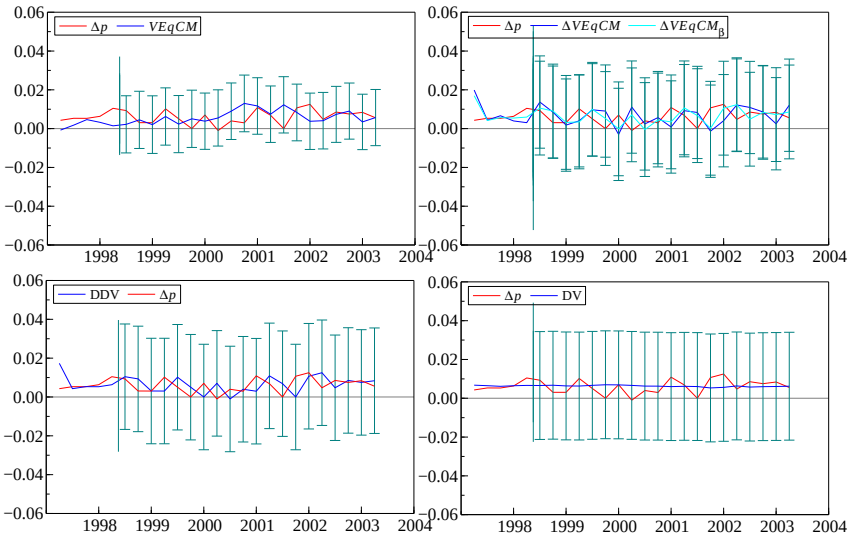

Figure 5.6: 1-step forecasts of quarterly UK inflation from the VEqCM, ΔVEqCM and ΔVEqCM_β, DDV and DV

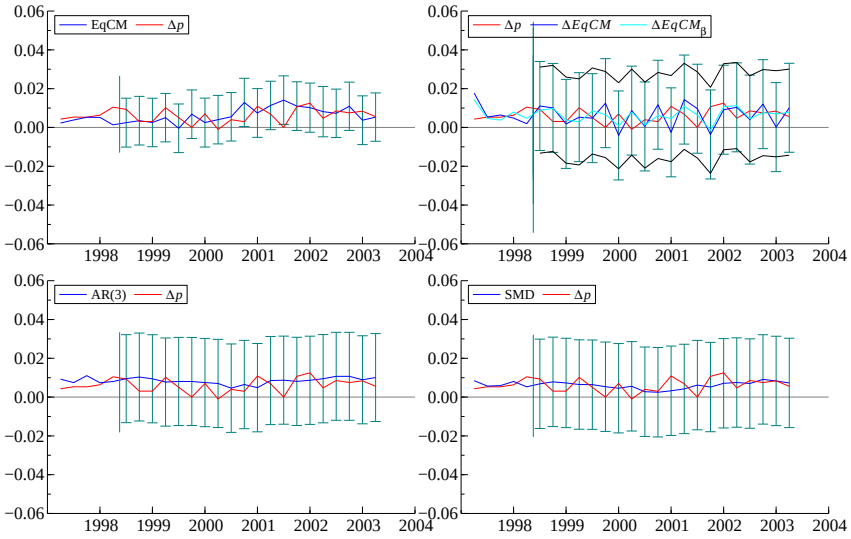

Figure 5.7: 1-step forecasts of quarterly inflation from the EqCM, ΔEqCM and ΔEqCM_β, AR(3) and SMD

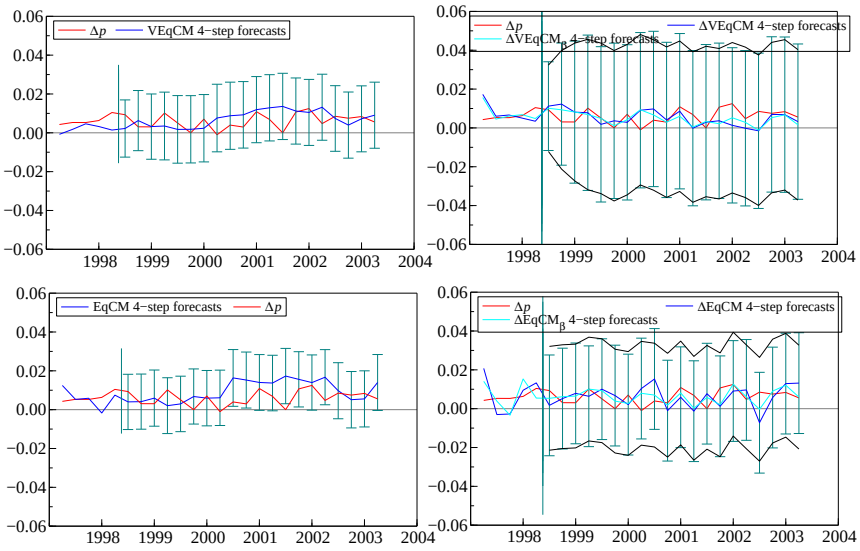

Figure 5.8: 4-step forecasts for quarterly inflation derived from the VEqCM, ΔVEqCM and ΔVEqCM$_\beta$, EqCM, ΔEqCM and ΔEqCM$_\beta$ models

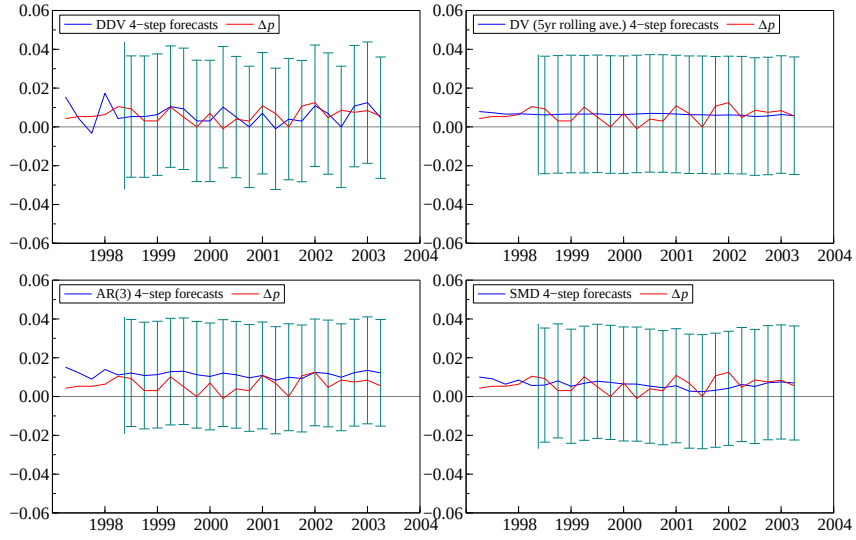

Figure 5.9: 4-step forecasts for quarterly inflation derived from the DDV, DV, AR(3) and SMD

Table 5.10: Summary of 20-period dynamic quarterly inflation forecasts

Dynamic Forecasts (20 periods)	*ME*	*MAE*	*RMSFE*
VEqCM	-0.183	0.400	0.480
AR(1) powered (1966q1-1998q2)	-1.187	1.187	1.254
AR(1) powered (1992q1-1998q2)	-0.069	0.324	0.387

Notes: Dynamic forecasts for 1998q3-2003q2. Figures reported as percentages.

for the VEqCM, against a powered up AR(1) model given by:

$$\Delta p_t = \alpha + \beta \Delta p_{t-1} + \epsilon_t \tag{5.42}$$

$$\Delta p_{T+h|T} = \sum_{i=0}^{h-1} \alpha\beta^i + \beta^h \Delta p_T. \tag{5.43}$$

The RMSFE of the VEqCM for 20 periods ahead of 0.48% is excellent, reflecting the relative stability of inflation over the forecast period and the lack of deterministic trend in the VEqCM model. Two AR(1) models are examined; the first is estimated over the full sample and the second is estimated from 1992 onwards. Estimating over the full sample period results in a RMSFE of 3.5 times that of the smaller sample, demonstrating the extent of structural change over the sample period. If the model is a good representation of the economy and the structure of the economy remains relatively unchanged, then forecast accuracy should decline as the forecast horizon increases because innovation errors accumulate and predictability falls. However, if models are mis-specified and unanticipated shifts occur, particularly in the deterministic terms, classical forecast theory does not hold. In this case, forecast failure can easily arise and it may be possible that forecast accuracy increases with the forecast horizon. Hence, even though inflation is I(0) over the forecast horizon, breaks may well be causing forecast failure, as demonstrated by the 20-step forecasts outperforming most 1-step and 4-step forecasts considered in Table 5.9.

5.3.4 Ranking of forecasting models

Having assessed the models on MSFE criterion, this section aims to rank the models in terms of the closest absolute forecast to actual inflation. Furthermore, the models are examined to establish whether there is any autocorrelation in which models perform best (and worst) over the forecast horizon, with a view to providing a more informed choice regarding which models to use. The models are also tested for how well they predict sign

changes in inflation, although the relative stability of inflation over the forecast horizon implies that this is less relevant than in periods when inflation is closer to an I(1) process.

On RMSFE and MAE criteria, the SMD and the DV are preferred over the 1-step horizon and the DV and pooled forecast perform best over the 4-step horizon. A question worth addressing is which of these models delivers the closest forecast at each forecast realisation. Table 5.11 ranks the top two forecasting models based on the absolute forecast error at every horizon. The worst model is also reported. Although the table is rather cumbersome to read, it is clear that no model systematically outperforms the other models over the forecast period, and likewise, no model systematically has the largest absolute errors. The magnitude of the errors is fairly similar across models and there is a lot of fluctuation with regard to which models deliver the smallest absolute errors. Forecast accuracy appears to be rather volatile, with forecasting models being ranked both 'best' and 'worst' over the forecast period. Most models tend to perform well for some quarters and poorly for others, and so we cannot draw conclusions regarding a systematic ranking.

This analysis is conditional on the sample examined. Inflation is relatively stable over the period and all the forecast errors are of a similar magnitude. During a period in which inflation is more volatile there may well be some systematic rankings. Identifying why particular models perform well would be the logical step in this case.

5.3.5 Forecasting correct sign changes

Another important question to address when examining forecasting models is whether the models correctly forecast signs and changes in signs. Over the forecast period quarterly inflation was positive, other than on one occasion when inflation was negative (2000q2), and on two occasions was zero (1999q4 and 2001q3). None of the models correctly forecast the negative inflation observation, although the inflation realisation was just -0.1%. Most models always predict positive inflation, other than the differenced EqCM and VEqCM models, and the DDV predicts one negative inflation observation in the quarter following 2000q2. Given the Bank of England's inflation target, predicting the sign of inflation is not an informative model criterion.

Table 5.11: Model rankings based on absolute forecast error

	1-step Forecasts			*4-step Forecasts*		
	1st	*2nd*	*Worst*	*1st*	*2nd*	*Worst*
1998q3	ΔEqCM	AR(3)	VEqCM	ΔVEqCM$_\beta$	ΔVEqCM	ΔEqCM
1998q4	VEqCM	EqCM	AR(3)	EqCM	ΔEqCM	ΔVEqCM
1991q1	DDV	DVEqCM$_\beta$	AR(3)	VEqCM	SMD	AR(3)
1999q2	AR(3)	SMD	ΔEqCM$_\beta$	ΔEqCM$_\beta$	DDV	EqCM
1999q3	Pool	ΔVEqCM	EqCM	ΔVEqCM$_\beta$	VEqCM	AR(3)
1999q4	DDV	ΔEqCM	ΔVEqCM	ΔVEqCM$_\beta$	DDV	AR(3)
2000q1	DV	AR(3)	ΔVEqCM	SMD	DV	ΔEqCM
2000q2	EqCM	SMD	ΔVEqCM	EqCM	SMD	AR(3)
2000q3	Pool	AR(3)	VEqCM	DDV	SMD	EqCM
2000q4	DDV	SMD	VEqCM	ΔVEqCM$_\beta$	ΔVEqCM	EqCM
2001q1	EqCM	VEqCM	DDV	AR(3)	VEqCM	SMD
2001q2	DV	ΔVEqCM	VEqCM	DV	AR(3)	VEqCM
2001q3	SMD	DV	EqCM	SMD	ΔVEqCM	EqCM
2001q4	EqCM	AR(3)	ΔEqCM$_\beta$	AR(3)	Pool	ΔEqCM
2002q1	ΔEqCM	DDV	$\frac{1}{4}\Delta_4 p_{t-1}$	AR(3)	ΔEqCM$_\beta$	ΔVEqCM
2002q2	SMD	DV	DDV	DV	ΔEqCM$_\beta$	EqCM
2002q3	ΔVEqCM$_\beta$	EqCM	ΔEqCM	EqCM	AR(3)	ΔEqCM
2002q4	ΔEqCM$_\beta$	ΔVEqCM$_\beta$	ΔVEqCM	Pool	SMD	AR(3)
2003q1	AR(3)	SMD	ΔVEqCM	SMD	Pool	AR(3)
2003q2	EqCM	DV	AR(3)	DV	ΔEqCM$_\beta$	EqCM

A more informative indication of the models relative forecast accuracy would be to examine how well the models forecast the direction of the change in inflation. Over the 20 forecast observations, quarterly inflation rose in 8 periods and fell in 12 periods. There is no systematic movement in inflation, as can be seen in Figure 5.10b, which records the first difference of quarterly inflation. Figures 5.10c and 5.10d record the percentage of observations in which each forecasting model correctly predicts the sign change in inflation. Both the 1-step and 4-step forecasts suggest that for these models, the chance of predicting a rise or fall in inflation is not sytematically correct, and indeed, is more often incorrect than correct; a systematic finding across almost all models. Again, we need to apply the caveat that the results are sample specific and may be due to the nature of inflation of over the forecast horizon.

Having looked at the forecasting models in terms of their rankings based on absolute errors and which models predict sign changes, we may be rather pessimistic over the conclusions. There is no systematic 'good' model and none seem to forecast sign changes correctly on average. However, we need to bear in mind the small magnitude of errors.

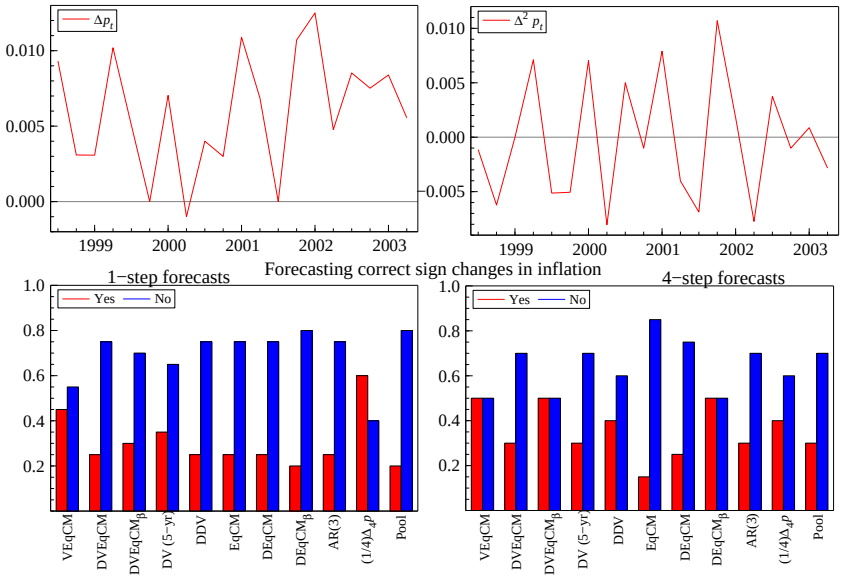

Figure 5.10: Quarterly inflation, the change in quarterly inflation, and the ability of the forecasting models to predict the sign change in inflation for the 1-step and 4-step forecasts

A key issue given the similarity of forecast errors between models is whether any of the models would have led to the Bank of England missing the inflation target band. Issues such as the time lag, forecasts conditional on the path of interest rates, the balance of uncertainties etc. also need to be addressed.

5.4 Models of annual inflation

Forecasting requires predictability, where a process y_t is defined as predictable with respect to an information set, $\mathcal{I}_{t-1}$, over $\mathcal{T}$ if $D_{y_t}(y_t|\mathcal{I}_{t-1}) \neq D_{y_t}(y_t)$ for $\forall t \in \mathcal{T}$: see Chapter 1 for a brief discussion. As noted, predictability is relative to the information used and so forecasting from a reduced, but proper, information set, $\mathcal{J}_{t-1} \subset \mathcal{I}_{t-1}$, will result in less accurate but unbiased predictions. Unpredictability is not invariant to the data frequency used and so temporal disaggregation cannot lower the predictability of y_t. This implies that as lower frequency data is a subset of higher frequency data, more accurate predictions should be obtained when forecasting annual inflation using quarterly data as opposed to annual data, although both forecasts should be unbiased on a proper information set.

In order to test the theory, we derive two models of annual inflation. We examine the

simple case where $y_t = f_t(\mathcal{I}_{t-1}) + \nu_t$, for a model in which annual inflation is forecast using a single-equation dynamic model based on *Gets* methodology. The first model we examine is the annual analogue of that on quarterly data, requiring information in fourth differences and lagged four periods only to emulate annual data. This is our lower frequency model. We then develop a model of annual inflation using quarterly data representing the higher frequency model, and examine the improvement in forecasting performance when the information set is increased. We use the time subscript τ to denote the annual frequency in order to distinguish between the quarterly and annual inflation models.

5.4.1 Annual analogue model of annual inflation

The initial model of $\Delta_4 p_\tau$ includes lags 4 to 7 of y^d, U^d, e_r, $(c^* - p)$, s, $(ppi - p)$, $(rent - p)$, $(imp - p)$, $(oil - p)$, $(n - p)$, R_l, $\Delta_4 ppi$, $\Delta_4 rent$, $\Delta_4 imp$, $\Delta_4 oil$, $\Delta_4 p$, $\Delta_4 c^*$, $\Delta_4 m4$, $\Delta_4 n$, $\Delta_4 R_s$, $\Delta_4 R_l$, $\Delta_4 pw$, $\Delta_4 e_\tau e_{r,\tau-j}$ and $\Delta_4 e_\tau e^2_{r,\tau-j}$ for $j = 4, 5$, and an intercept and trend. Five indicator variables are included given by:

$$D_t = [I_{72:4}, I_{73:2}, I_{74:1}, I_{84,1}, I_{84:2}]',$$

and a year long dummy variable is included for 1979q3-1980q2. To derive a model of annual inflation in annual analogues using quarterly data autocorrelation must be corrected for. To still use the *Gets* framework embodied in PcGets, the standard errors need to be adjusted as least squares will be inefficient. The first order autocorrelation is of magnitude 0.7, with highly significant second and third order autocorrelation. As a rough guide, we can adjust the t-values that are selected in PcGets by a factor $\sqrt{\left(\frac{1+\rho}{1-\rho}\right)}$. Hence, we shall initially retain variables with a t-statistic greater than 4.76.[15] To refine the selection using a more rigorous adjustment for autocorrelation, the model is then estimated in PcGive and tested down using HACSEs.

The resulting model is reported in (5.45), with HACSEs reported in parentheses. The model fails autocorrelation, as expected, and the RESET test of model specification,

[15] This can be done by adjusting the probabilities for the t-tests in the expert-users strategy option in PcGets.

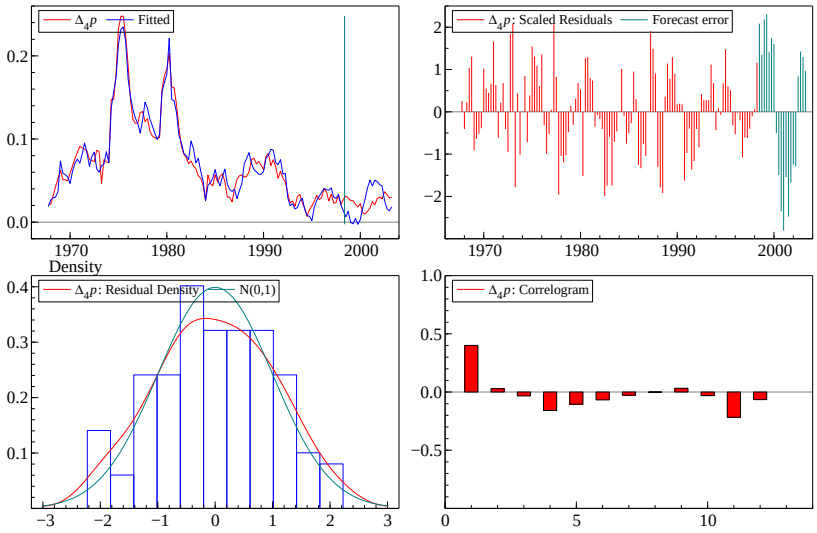

Figure 5.11: Model of annual inflation using quarterly data in annual analogues; model fit, residuals, their density and correlogram

which may be an indication of some nonlinearity. The model also fails the forecast Chow test, but out-of-sample fit is not a criterion on which to base in-sample model selection as this would bias the forecasting results. The model contains most theories of inflation found in Hendry (2001), and the mark-up is given in (5.44):

$$\pi^*_\tau = p_\tau - 0.58c^*_\tau - 0.34pw_{£,\tau} - 0.07ppi_\tau + 4.12. \tag{5.44}$$

The corresponding graphics including the model fit, residuals, density, and correlogram are given in Figure 5.11, and the recursive graphics are recorded in Figure 5.12.

$$\begin{aligned}
\Delta_4 p_\tau &= \underset{(0.092)}{-0.333}\Delta_4 p_{\tau-5} + \underset{(0.067)}{0.801} y^d_{\tau-4} + \underset{(0.071)}{0.258}\Delta_4 c^*_{\tau-5} - \underset{(0.062)}{0.341} U^d_{\tau-6} \\
&\quad + \underset{(0.040)}{0.092}\Delta_4 rent_{\tau-6} + \underset{(0.008)}{0.035}\Delta_4 oil_{\tau-4} + \underset{(0.146)}{0.640}\Delta_4 pw_{\tau-4} - \underset{(0.043)}{0.325}\pi^*_{\tau-4} \\
&\quad - \underset{(0.005)}{0.058} I_{74:1} - \underset{(0.004)}{0.033} I_{84:1} + \underset{(0.005)}{0.043} D_{79:3,80:2} + \underset{(0.005)}{0.026} \qquad (5.45) \\
\mathsf{R}^2 &= 0.951 \quad \widehat{\sigma} = 1.202\% \quad SIC = -8.476 \\
\mathsf{F}_{\mathsf{ar}}(5,106) &= 6.599^{**} \quad \mathsf{F}_{\mathsf{arch}}(4,103) = 1.792 \quad \mathsf{F}_{\mathsf{het}}(19,91) = 1.328 \quad \chi^2_{\mathsf{nd}}(2) = 1.440 \\
\mathsf{F}_{\mathsf{reset}}(1,110) &= 19.169^{**} \quad \mathsf{F}_{\mathsf{Chow}}(20,111) = 2.469^{**} \quad T = 1967q4 - 1998q2.
\end{aligned}$$

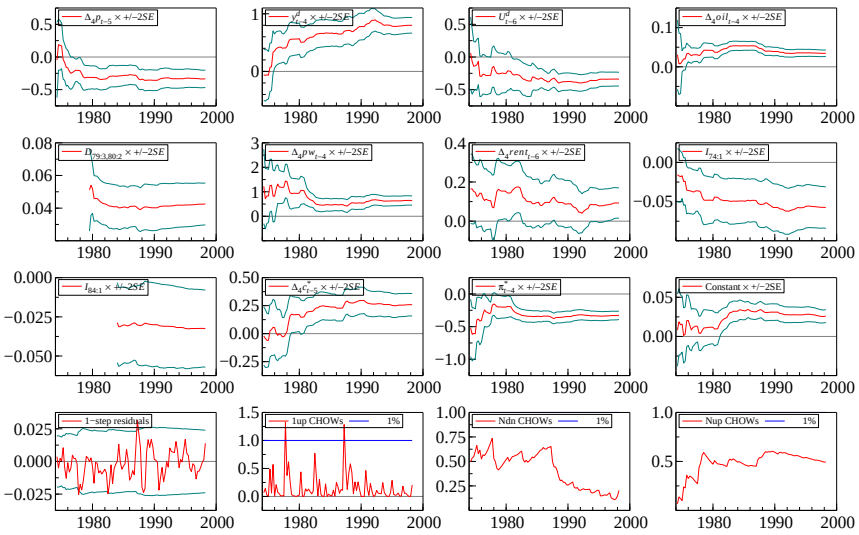

Figure 5.12: Model of annual inflation using quarterly data in annual analogues; recursive coefficients with $\pm 2\sigma$, 1-step residuals and constancy tests

5.4.2 Quarterly model of annual inflation

A model of annual inflation using quarterly data is developed in order to compare the forecasts with (5.45). The model is selected using the same methodology in which first differences are included in the GUM to reflect the higher frequency data. Again, we only use lags dated $\tau-4$ and previous in order to forecast 1-year ahead. The model is reported in (5.46), with standard errors adjusted for autocorrelation and heteroskedasticity in parentheses.

$$\begin{aligned}
\Delta_4 p_\tau &= \underset{(0.056)}{0.711} y^d_{\tau-4} + \underset{(0.080)}{0.609}\Delta c^*_{\tau-4} - \underset{(0.048)}{0.405} U^d_{\tau-7} + \underset{(0.005)}{0.021}\Delta oil_{\tau-7} \\
&\quad + \underset{(0.043)}{0.258}\Delta ppi_{\tau-4} + \underset{(0.323)}{1.528}\Delta pw_{\tau-5} - \underset{(0.037)}{0.279}\pi^*_{\tau-5} - \underset{(0.003)}{0.022} I_{73:2} \\
&\quad - \underset{(0.006)}{0.041} I_{74:1} + \underset{(0.003)}{0.060} D_{79:3,80:2} + \underset{(0.006)}{0.079} D_{75:2,75:3} + \underset{(0.003)}{0.032} \qquad (5.46) \\
\mathsf{R}^2 &= 0.952 \quad \widehat{\sigma} = 1.189\% \quad SIC = -8.497 \\
\mathsf{F}_{\mathsf{ar}}(5,106) &= 1.531 \quad \mathsf{F}_{\mathsf{arch}}(4,103) = 0.191 \quad \mathsf{F}_{\mathsf{het}}(18,92) = 1.098 \quad \chi^2_{\mathsf{nd}}(2) = 5.862 \\
\mathsf{F}_{\mathsf{reset}}(1,110) &= 4.766^* \quad \mathsf{F}_{\mathsf{Chow}}(20,110) = 1.513 \quad T = 1967q4 - 1998q2.
\end{aligned}$$

There is very little improvement in fit moving to the higher frequency data, with an equation standard error of 1.19% as opposed to 1.2% for the annual model. The corresponding graphics including the model fit, residuals, density, and correlogram are given

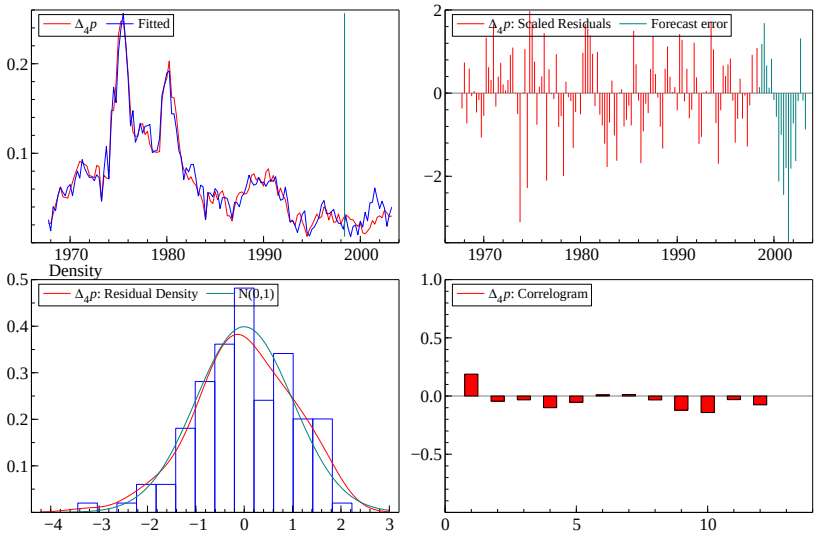

Figure 5.13: Model of annual inflation using quarterly data; model fit, residuals, their density and correlogram

in Figure 5.13, and the recursive graphics are recorded in Figure 5.14.

5.4.3 Quarterly inflation model used to forecast 1-year ahead inflation

An alternative rule that we investigate is based on deriving a model of annual inflation using quarterly data and fixing the estimated coefficients to forecast 4-steps ahead, as opposed to 1-step ahead. If the DGP is given as:

$$\Delta_4 \mathbf{x}_t = \gamma + \alpha \left(\beta' \mathbf{x}_{t-1} - \mu \right) + \nu_t, \tag{5.47}$$

we use the forecasting rule given by:

$$\widehat{\Delta_4 \mathbf{x}}_{T+4|T} = \widehat{\gamma} + \widehat{\alpha} \left(\widehat{\beta}' \mathbf{x}_{T-1} - \widehat{\mu} \right). \tag{5.48}$$

While this is a mis-specified model, we can interpret the forecasting rule as:

$$\Delta_4 \mathbf{x}_{T+4|T} = \Delta_4 \mathbf{x}_T - \widehat{\nu}_t, \tag{5.49}$$

and so the forecasting rule is the DDV when there are no breaks, excluding the estimated error term. As well as protecting against breaks via the DDV component, all available information up to time T is used to develop the dominant, congruent, in-sample model. To produce a direct one year ahead forecast, quarterly information is lost because the

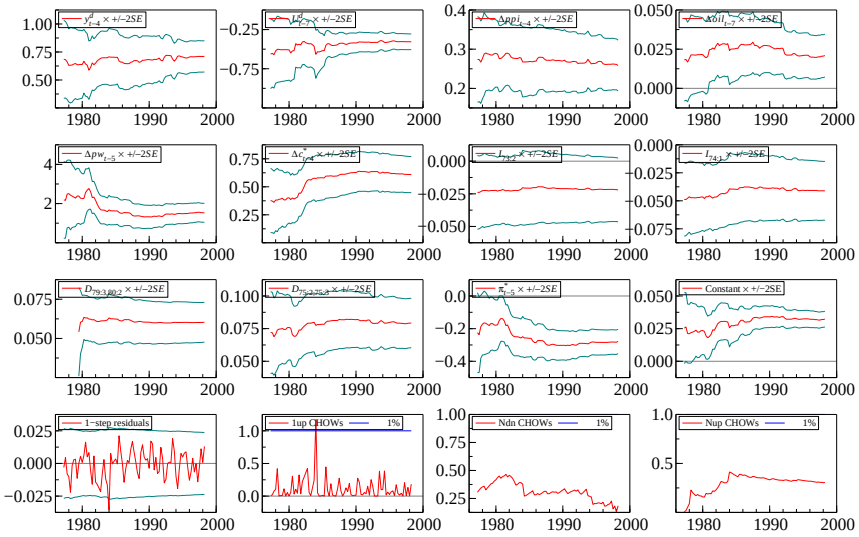

Figure 5.14: Model of annual inflation using quarterly data; recursive coefficients with $\pm 2\sigma$, 1-step residuals and constancy tests

in-sample model is based on annual data. Developing a quarterly model and using the resulting coefficients lagged by one year requires the implicit assumption that the exogenous variables have the same impact lagged one quarter as they do lagged one year.

We derive a model of annual inflation by imposing a coefficient of unity on $\Delta_3 p_{\tau-1}$ to model quarterly inflation. The selection process retained the lagged dependent variables, and a test for the restriction $\Delta p_{\tau-4} = \Delta p_{\tau-5} = \Delta p_{\tau-6} = 1$ is accepted; $F(1,110) = 0.438\,[0.509]$. Having derived the quarterly model of annual inflation, the coefficients are fixed and the model is re-formulated on the variables lagged one year. The resulting model is given in (5.50):

$$\begin{aligned}
\Delta_4 p_\tau &= 1.000\Delta_3 p_{\tau-4} + 0.208 y^d_{\tau-4} + 0.194\Delta c^*_{\tau-4} - 0.134 U^d_{\tau-4} \\
&\quad +0.109\Delta imp_{\tau-5} + 0.414\Delta pw_{\tau-7} - 0.021\pi^*_{\tau-5} + 0.018 I_{72:4} \\
&\quad -0.046 I_{73:2} - 0.030 I_{74:1} + 0.045 I_{79:3} + 0.018 I_{84:2} + 0.007 \qquad (5.50) \\
\widehat{\sigma} &= 3.240\% \quad SIC = -6.859 \quad \mathsf{F}_{\mathsf{ar}}(5,118) = 49.075^{**} \quad \mathsf{F}_{\mathsf{het}}(19,103) = 9.973^{**} \\
\chi^2_{\mathsf{nd}}(2) &= 7.774^{**} \quad \mathsf{F}_{\mathsf{Chow}}(20,123) = 0.0386 \quad T = 1967q4 - 1998q2.
\end{aligned}$$

The model is a poor fit and is clearly mis-specified, failing many diagnostics, although this is not indicative criteria for the model's out-of-sample performance.

Table 5.12: Forecast performances of models for 4-quarter ahead annual inflation

Model	*ME*	*MAE*	*RMSFE*
Ann EqCM	0.11	1.96	2.08
Qu EqCM	-0.53	1.26	1.58
Qu ΔEqCM$_\beta$	0.11	0.87	1.02
DDV$_{\text{1yr}}$	0.12	0.79	0.93
AR(4) (lags 4-7)	-1.62	1.62	1.84
AR(1) (powered up)	-0.71	0.91	1.11
Qu EqCM$_{\text{in-sample}}$	**0.02**	**0.53**	**0.64**
Pool	-0.34	0.60	0.77

Notes: Figures reported as percentages, with the best forecasting performance highlighted in **bold**.

5.5 Annual inflation forecasts

Table 5.12 reports the one year ahead ME, MAE, and RMSFE of the annual inflation forecasting models, for the period 1998q3-2003q2. The models include the annual analogue derived in (5.45), the quarterly inflation model derived in (5.46), and the differenced quarterly model excluding double differenced terms. Other models include a DDV forecasting one year ahead given by $\widehat{\Delta_4 \mathbf{x}}_{T+4|T} = \Delta_4 \mathbf{x}_T$, an AR(4) model that gives direct forecasts as the regressors are lagged four periods, an AR(1) model in which the forecasts are derived by powering up the lagged coefficient, and equation (5.50), in which the model is a quarterly EqCM in-sample model, but the coefficients are fixed for the 4-quarter lagged regressors to deliver direct 4-step ahead forecasts.

Figure 5.15 records the forecasts from the models outlined above. The quarterly model forecasts outperform those from the annual model despite a substantial upward bias, suggesting that there are gains to be made from moving to higher frequency data, notwithstanding very little improvement in in-sample fit. Time disaggregation is beneficial in this context. However, the concern of increased noise due to measurement errors may be a problem with higher frequency data, and caution must be applied when choosing the optimal data frequency. The AR(4) model has a substantial bias due to the higher in-sample mean. Both the DDV and quarterly EqCM in-sample model perform well, with a RMSFE of 0.93% and 0.64% respectively. The model based on (5.50) is likely to be capturing the relative stability of inflation over the forecast horizon, but the method would also be robust to large structural breaks in inflation. Pooling is shown to

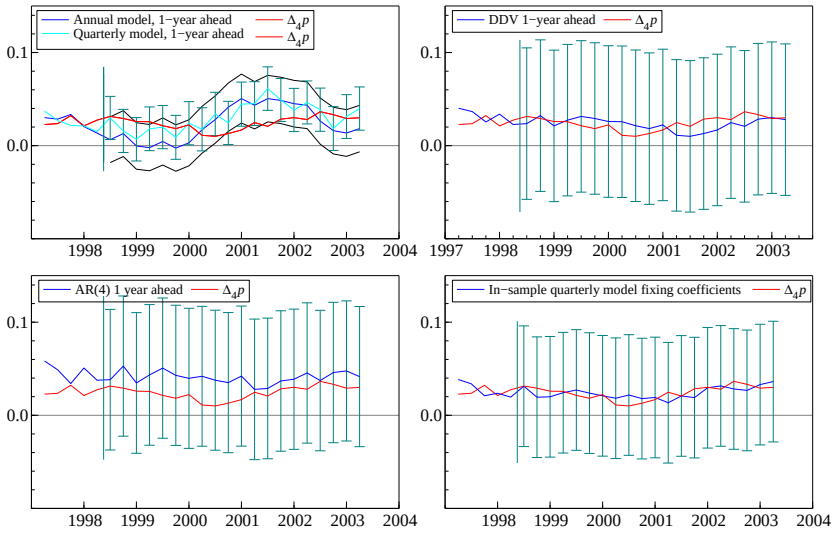

Figure 5.15: 1-year ahead forecasts of annual inflation from the annual EqCM and the quarterly EqCM, the DDV, the AR(4) and the in-sample quarterly model

be beneficial, indicating that no model encompasses the others.

5.6 Conclusion

A VEqCM model of quarterly inflation is developed, conditioning on excess demand. We undertake an extensive analysis of the properties of the cumulated output gap as concerns about its reliability restrain us from including it in the cointegrating space. As the critical values are incorrect for the model estimated, we make an informed decision as to the cointegrating rank based on all available evidence. Imposing a rank of one results in a VEqCM with the cointegrating vector capturing the mark-up. A reduction to a more parsimonious VEqCM is undertaken and this, along with the single-equation model developed in Chapter 4, forms the basis for the substantive component of the chapter, examining the forecast performance of the econometric models in relation to the forecast performance of a variety of robust forecasting devices.

The analysis has highlighted the importance of deterministic terms when forecasting, and particular attention should be paid to $\boldsymbol{\mu}$ and $\boldsymbol{\gamma}$. Shifts in the mean of the cointegrating vector or the unconditional growth rate of the system will lead to forecast failure. However, it is often difficult to identify where breaks are occurring in the data. Pooling is shown to be successful in many situations, and this could be seen as an 'insurance

policy' type forecast rather than selecting a particular forecasting method *ex ante*.

We find that there are benefits to disaggregation. In this chapter we have looked at data frequency but Hendry and Hubrich (2006) also look at disaggregation across variables and space. Moving to higher frequency data should enable breaks to be picked up sooner and so the forecasting method could switch to the robust forecasting device faster.

The empirical example of UK quarterly and annual inflation demonstrates the success of adaptive forecasting devices outlined in the Clements and Hendry (1998b, 1999) theory of forecasting. While there are no obvious structural breaks at the forecast horizon, there is a clear regime shift throughout the 1990s in which inflation is reduced to levels previously seen in the 1950s. This follows the fall out of the ERM and the subsequent switch to inflation targeting. The move to Central Bank independence in 1997 has brought about low and stable inflation, and whether this is permanent or transitory, robust forecasting devices are picking up this behaviour, whereas congruent in-sample models are still correcting to 'the old' equilibrium. Hence, the application of the forecasting theory to inflation does yield the results predicted by the theory.

Epilogue

The dual issues of model selection and forecasting are controversial in econometrics. The debates over data-mining, data-based model selection, and measurement without theory still pervade the model selection literature, and the irrefutable evidence exposing the forecast failure of many econometric models undermines much of the forecasting literature. However, the situation is not so bleak. Huge advances in both fields have led to new theories, techniques and practices that overcome many of the criticisms previously advanced. Automatic econometric model selection has come to the fore with the development of PcGets, which selects models using a general-to-specific search strategy in which many paths are searched. The properties of the algorithm are excellent under a wide range of linear states of nature, and the thesis has further established its properties under nonlinearity. The Clements and Hendry (1998b, 1999) forecasting theory provides an explanation as to why forecast failure is pervasive in econometrics: unanticipated structural breaks occur frequently in macroeconomic time-series and these are disastrous for the equilibrium correction class of models, leading to the development of forecasting devices that are robust to structural breaks. This thesis has built on the new research, investigating the selection of forecasting models and nonlinear models using automatic model selection tools, and undertaking a forecast competition in which econometric models are pitted against robust forecasting devices. We have not assessed the forecasting performance of nonlinear inflation models, and this is a topic of future research.

Chapter 2 assessed the automatic model selection algorithms, PcGets and RETINA. The programs have different structures and objectives and should be viewed as complementary procedures that each work optimally under different settings. PcGets selects a congruent, undominated model of the phenomena of interest, whereas RETINA selects a parsimonious set of regressors that have predictive ability. We compared the two programs on cross-section and time-series data with promising results for both algorithms:

the ranking of the algorithms on both in-sample and forecasting performance was not clear cut. The predominant feature of RETINA is the parsimony it can achieve from a highly over-parameterised GUM, and this is due to the use of three disjoint sub-samples in the selection algorithm. The trade-off is that its selection criteria are very stringent, which can lead to a failure to retain highly significant variables. The empirical applications highlighted the need to check for extreme observations, as the nonlinear functions generated by RETINA may simply be reflecting a few outliers. A Monte Carlo analysis of the selection properties of the algorithms for nonlinear models was also undertaken. The evidence is encouraging: PcGets rejection frequencies do not differ substantially if orthogonal nonlinear functions are included. Furthermore, the null rejection frequency of RETINA is found to be much tighter for large GUMs. The evidence for non-orthogonal models is less satisfactory as the null rejection frequency is over-sized and the non-null rejection frequency is too small, both for PcGets and RETINA. However taking deviations from means to result in a near-orthogonal specification dramatically improves the selection properties of both algorithms, and the chapter concludes by observing that nonlinearity can be tested for at low cost using automatic model selection algorithms.

Chapter 3 developed a strategy for selecting nonlinear models within the *Gets* framework embedded in PcGets. The strategy tests for evidence of nonlinearity within the initial batch of diagnostic tests undertaken to establish congruency of the GUM. If a linear functional form is rejected, a GUM is generated in which polynomial functions of the postulated variables are included, where the functions are double de-meaned prior to inclusion to ensure a near orthogonal representation. The specific model is then selected concurrently with the application of the indicator saturation technique to identify outliers. This avoids retention of too many nonlinear functions that are proxying indicators for extreme observations. The selection stage is undertaken at tighter significance levels for the nonlinear functions to avoid excess retention of irrelevant nonlinear functions which may be particularly detrimental for forecasting. A multi-stage procedure is recommended, ensuring that nonlinearity is retained throughout the search procedure if the index test found evidence of nonlinearity in the GUM. This will result in an undominated,

congruent, nonlinear, specific model.

The chapter has established the properties of the nonlinearity test, demonstrating that it has power over alternative tests when there are a large number of regressors in the GUM, or when the degree of collinearity between regressors is high. Furthermore, the use of polynomials was motivated by their ability to approximate an LSTR model, which enables generality of the GUM functional form to be retained. The establishment of simple de-meaning rules to remove collinearity is of fundamental importance: many applied papers include quadratic and cubic terms in empirical models with no adjustment, generating unwanted collinearity. The feasibility of automatically selecting nonlinear models has been established, with the suprising result that even though we remain agnostic regarding the functional form of the potential nonlinearity, we can retain the selection properties of the PcGets algorithm established under linearity. The success of the proposed selection algorithm is due to the synthesis of all components of the algorithm, which is a vital contribution to the model selection literature.

Chapter 4 built a model of UK inflation based on many theories of inflation, including unit labour costs, import prices, producer prices, exchange rates, foreign prices, excess demand for labour and excess demand for goods and services, using PcGets. The development of the inflation model required a measure of the output gap. A production function approach was initially adopted, and attempts to account for cyclicality proved successful in developing a measure of TFP. Concerns over measurement errors led to the adoption of a dynamic production function model in which TFP was modelled as a random walk with drift. While this had the advantage of allowing for adjustments to equilibrium over time, the measure of TFP soaked up much of the short-run volatility in output, resulting in a smaller gap. The final measure imposed TFP to follow a segmented linear trend, generating larger cycles in the output gap.

The inflation model captures the large swings in inflation over the period, which includes many regime shifts, with constant parameters and few indicators. The evidence strengthens the argument that there is no single cause of inflation, which led to an investigation as to whether a generic 'business cycle' factor could explain inflation.

Principal components of UK inflation were derived based on a moderately large database of explanatory variables. The evidence indicates that business cycle pressures are not sufficient to explain inflation alone, and furthermore, they add little explanatory power to the model based on individual explanations of inflation.

Chapter 5 examined the forecasting performance of econometric models, in comparison to various 'robust' forecasting devices. Initially a VEqCM model of inflation was developed using partial cointegration analysis, in which the system comprised the domestic price level, unit labour costs, producer prices, world prices and the nominal exchange rate, and conditioned on the output gap. Exclusion of the cumulated output gap was justified on the basis of pernicious measurement error in the gap estimates. While there was some evidence of autocorrelation in the system, the inflation model was comparable to that of the single-equation model. The results of the forecasting competition were inconclusive: no model performed systematically better than any other model. However, the robust forecasting devices were comparable to the econometric models, demonstrating just how difficult it is to 'beat' simple models such as random walks. Pooling was shown to be successful in many situations. Disaggregation proved to be beneficial, providing support for the use of higher frequency data. The forecast period was relatively quiescent, which proved to be an interesting period on which to test the robust forecasting devices. The theory argues that robust forecasting devices should perform well when there are structural breaks. However, we have shown that they also perform well in quiet periods, providing further support for alternative methods of model selection for forecasting models.

While this thesis has covered much ground, there are many fruitful areas on which to build. The most pressing is the nonlinear approximations, in which substantial theoretical and Monte Carlo analysis is needed to establish the ability of various approximations to capture potential nonlinear functional forms. An investigation into orthogonal polynomials, Fourier series and asymptotic series was undertaken, but further work on confluent hypergeometric series may prove useful. As the LSTR nests many nonlinear specifications, we are reasonably confident that an ability to capture the properties of an LSTR

implies the approximation will perform well for a variety of other nonlinear specifications. More extensive Monte Carlos are also required to establish the power properties of the nonlinearity test under a much wider range of scenarios. The analysis in Chapter 3 examined the power of the index test under the most challenging settings, and it is predicted that the test will perform even better under alternative scenarios, such as the inclusion of additional nonlinear functions. Selection on the GUM prior to application of the nonlinearity test may improve the power properties by reducing the degrees of freedom, and this is a further area of investigation. Translation of the proposed nonlinear model selection procedure into an automatic algorithm is, of course, the primary objective, and a comprehensive Monte Carlo analysis across a wide range of states of nature is required to calibrate the super-conservative strategy. Not only is a nonlinear capability for PcGets feasible, but it is an essential extension that would broaden its scope in keeping with a general-to-specific philosophy.

On the forecasting side, forecast comparison exercises are necessarily sample specific. However, given the development of nonlinear model selection tools, a natural progression would be to establish the forecasting properties of nonlinear models selected by the proposed algorithm. A key question to address is whether nonlinear models can be made robust to structural breaks. The thesis has laid the groundwork for developing forecasting models using automatic model selection tools, and further progress in this field is inevitable.

The framework outlined in the Introduction asserted that the thesis was broadly divided into two related parts; model specification and non-stationarity. Chapters 2 and 3 addressed model specification from an automatic model selection viewpoint and Chapters 4 and 5 focused on non-stationarity, building a constant parameter model of inflation over a period with many breaks and regime changes, and forecasting using techniques that are robust to breaks. We anticipate the synthesis of Parts I and II of the thesis will yield fruitful results, with an abundance of empirical applications lending themselves to such an analysis.

Bibliography

Abadir, K. M. (1999). An introduction to hypergeometric functions for economists. *Econometric Reviews*, **18**, 287–330.

Abramowitz, M., and Stegun, I. A. (eds.)(1972). *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables,* 9th edn. New York: Dover.

Akaike, A. (1973). Information theory and an extension of the maximum likelihood principle. In Petrov, B. N., and Csaki, F. L. (eds.), *Second International Symposium of Information Theory*, pp. 267–281. Budapest: Akademiai Kiado.

Amemiya, T. (1980). Selection of regressors. *International Economic Review*, **21**, 331–354.

Amemiya, T. (1981). Qualitative response models: A survey. *Journal of Economic Literature*, **19**, 1483–1536.

Anderson, G. A., and Bollerslev, T. (1998). Answering the skeptics: Yes, standard volatility models do provide accurate forecasts. *International Economic Review*, **39**, 885–905.

Anderson, T. W. (1994). *An Introduction to Multivariate Statistical Analysis.* Wiley Classics Library edn. New York: John Wiley & Sons.

Andrews, D. W. K. (1991). Heteroskedasticity and autocorrelation consistent covariance matrix estimation. *Econometrica*, **59**, 817–858.

Andrews, M. J., Minford, A. P. L., and Riley, J. (1996). On comparing macroeconomic forecasts using forecast encompassing tests. *Oxford Bulletin of Economics and Statistics*, **58**, 279–305.

Andrews, R. L. (1994). Forecasting performance of structural time series models. *Journal of Business and Economic Statistics*, **12**, 129–133.

Angstenberger, J. (1996). Prediction of the S&P 500 index with neural networks. In Taylor, J. G. (ed.), *Neural Networks and Their Applications*, pp. 143–152. Chichester: John Wiley & Sons.

Ashley, R., Patterson, D., and Hinich, M. (1986). A diagnostic test for nonlinear serial dependence in time series fitting errors. *Journal of Time Series Analysis*, **7**, 165–178.

Bai, J., and Perron, P. (1998). Estimating and testing linear models with multiple structural changes. *Econometrica*, **66**, 47–78.

Baillie, R. T., Bollerslev, T., and Mikkelsen, H.-O. (1996). Fractionally integrated generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, **74**, 3–30.

Ball, L., and Mankiw, N. G. (2002). The NAIRU in theory and practice. *Journal of Economic Perspectives*, **16**, 115–136.

Bårdsen, G., and Fisher, P. G. (1999). Economic theory and econometric dynamics in modelling wages and prices in the United Kingdom. *Empirical Economics*, **24**, 483–507.

Bårdsen, G., Fisher, P. G., and Nymoen, R. (1998). Business cycles: Real facts or fallacies?. In Strøm, S. (ed.), *Econometrics and Economic Theory in the 20th Century*, Econometric Society Monograph 32, Ch. 16. Cambridge: Cambridge University Press.

Bårdsen, G., Hurn, S., and McHugh, Z. (2003a). A smooth-transition model of the Australian unemployment rate. Working paper No. 1002, Department of Economics, Norwegian University of Science and Technology.

Bårdsen, G., Jansen, E. S., and Nymoen, R. (2003b). Econometric inflation targeting. *Econometrics Journal*, **6**, 429–460.

Bates, J. M., and Granger, C. W. J. (1969). The combination of forecasts. *Operations Research Quarterly*, **20**, 451–468.

Batini, N., and Greenslade, J. V. (2006). Measuring the UK short-run NAIRU. *Oxford Economic Papers*, **58**, 28–49.

Batini, N., Jackson, B., and Nickell, S. (2000). Inflation dynamics and the labour share in the UK. Discussion paper No. 2., Bank of England External MPC Unit.

Baxter, M., and King, R. G. (1999). Measuring business cycles. Approximate band-pass filters for economic time series. *The Review of Economics and Statistics*, **81**, 575–593.

Bera, A. K., and Higgins, H. L. (1993). A survey of ARCH models: Properties, estimation and testing. *Journal of Economic Surveys*, **7**, 305–362.

Berk, K. N. (1978). Comparing subset regression procedures. *Technometrics*, **20**, 1–6.

Berkowitz, J. (2001). Testing density forecasts, with applications to risk management. *Journal of Business and Economic Statistics*, **19**, 465–474.

Bernanke, B. S., and Boivin, J. (2003). Monetary policy in a data-rich environment. *Journal of Monetary Economics*, **50**, 525–546.

Beveridge, S., and Nelson, C. R. (1981). A new approach to decomposition of economic time series into permanent and transitory components with particular attention to measurement of the business cycle. *Journal of Monetary Economics*, **7**, 151–174.

Bhansali, R. J. (2002). Multi-step forecasting. in Clements and Hendry (2002a), pp. 206–221.

Blanchard, O. J., and Quah, D. (1989). The dynamic effect of aggregate demand and supply disturbances. *American Economic Review*, **79**, 655–673.

Blanchard, O. J., and Summers, L. H. (1988). Beyond the natural rate hypothesis. *American Economic Review*, **78**, 182–187.

Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, **31**, 307–327.

Bollerslev, T., Chou, R. Y., and Kroner, K. F. (1992). ARCH modelling in finance: A review of the theory and empirical evidence. *Journal of Econometrics*, **52**, 5–59.

Bolt, W., and van Els, P. J. A. (2000). Output gap and inflation in the EU. Staff report No. 44, De Nederlandsche Bank, Netherlands.

Bontemps, C., and Mizon, G. E. (2003). Congruence and encompassing. in Stigum (2003), pp. 354–378.

Bowdler, C., and Jansen, E. S. (2004). A mark-up model of inflation for the Euro area. Working paper No. 306, European Central Bank.

Box, G. E. P., and Jenkins, G. M. (1970). *Time Series Analysis, Forecasting and Control*. San Francisco: Holden Day.

Box, G. E. P., and Newbold, P. (1971). Some comments on a paper of Coen, Gomme, and Kendall. *Journal of the Royal Statistical Society, A*, **134**, 229–240.

Box, G. E. P., and Tiao, G. C. (1976). Comparison of forecast and actuality. *Applied Statistics*, **25**, 195–200.

Breiman, L. (1995). Better subset regression using the nonnegative garrote. *Technometrics*, **37**, 373–384.

Breusch, T. S. (1990). Simplified extreme bounds. in Granger (1990), pp. 72–81.

Breusch, T. S., and Pagan, A. R. (1980). The lagrange multiplier test and its applications to model specification in econometrics. *Review of Economic Studies*, **47**, 239–253.

Britton, E., Cutler, J., and Wardlow, A. (1999). The banks use of survey data. *Bank of England Quarterly Bulletin*, **39**, 177–182.

Britton, E., Fisher, P., and Whitely, J. (1998). Inflation report projections: Understanding the fan chart. *Bank of England Quarterly Bulletin*, **38**, 30–37.

Brock, W. A., Dechert, W. D., and Scheinkman, J. A. (1987). A test for independence based on the correlation dimension. SSRI Working Paper No. 8702, Department of Economics, University of Wisconsin.

Brooks, C. (1997). Linear and nonlinear (non-) forecastability of daily sterling exchange rates. *Journal of Forecasting*, **16**, 125–145.

Camba-Mendez, G., Kapetanios, G., Weale, M., and Smith, R. (2002). The forecasting performance of the OECD composite leading indicators for France, Germany, Italy, and the UK. in Clements and Hendry (2002a), pp. 386–408.

Campos, J., and Ericsson, N. R. (1999). Constructive data mining: Modeling consumers' expenditure in Venezuela. *Econometrics Journal*, **2**, 226–240.

Campos, J., Ericsson, N. R., and Hendry, D. F. (eds.)(2005). *General-to-specific Modelling*. International Library of Critical Writings in Econometrics. Cheltenham: Edward Elgar.

Campos, J., Hendry, D. F., and Krolzig, H.-M. (2003). Consistent model selection by an automatic *Gets* approach. *Oxford Bulletin of Economics and Statistics*, **65**, 803–819.

Cattell, R. B. (1966). The Scree test for the number of factors. *Multivariate Behavioural Research*, **1**, 245–276.

Chan, K. S., and Tong, H. (1986). On estimating thresholds in autoregressive models. *Journal of Time Series Analysis*, **7**, 179–194.

Chatfield, C. (1993). Calculating interval forecasts. *Journal of Business and Economic Statistics*, **11**, 121–135.

Chevillon, G., and Hendry, D. F. (2005). Non-parametric direct multi-step estimation for forecasting economic processes. *International Journal of Forecasting*, **21**, 201–218.

Chong, Y. Y., and Hendry, D. F. (1986). Econometric evaluation of linear macroeconomic models. *Review of Economic Studies*, **53**, 671–690. Reprinted in Granger (1990), Ch.17.

Chow, G. C. (1960). Tests of equality between sets of coefficients in two linear regressions. *Econometrica*, **28**, 591–605.

Chow, G. C. (1981a). A comparison of the information and posterior probability criteria for model selection. *Journal of Econometrics*, **16**, 21–33.

Chow, G. C. (1981b). Selection of econometric models by the information criteria. In Charatsis, E. G. (ed.), *Proceedings of the Econometric Society European Meeting 1979*, Ch. 8. Amsterdam: North Holland.

Chow, G. C. (1984). Random and changing coefficient models. in Griliches and Intriligator (1984), Ch. 21.

Christ, C. F. (1966). *Econometric Models and Methods*. New York: John Wiley & Sons.

Christoffersen, P. F., and Diebold, F. X. (2000). How relevant is volatility forecasting for financial risk management?. *Review of Economics and Statistics*, **82**, 12–22.

Clemen, R. T. (1989). Combining forecasts: A review and annotated bibliography. *International Journal of Forecasting*, **5**, 559–583.

Clements, M. P. (2005). *Evaluating Econometric Forecasts of Economic and Financial Variables*. UK: Palgrave Macmillan.

Clements, M. P., Franses, P. H., Smith, J., and van Dijk, D. (2003). On SETAR non-linearity and forecasting. *Journal of Forecasting*, **22**, 359–375.

Clements, M. P., Franses, P. H., and Swanson, N. R. (2004). Forecasting economic and financial time series with non-linear models. *International Journal of Forecasting*, **20**, 169–183.

Clements, M. P., and Galvão, A. B. (2004). A comparison of tests of non-linear cointegration with an application to the predictability of US interest rates using the term structure. *International Journal of Forecasting*, **20**, 219–236.

Clements, M. P., and Hendry, D. F. (1993a). On the limitations of comparing mean squared forecast errors. *Journal of Forecasting*, **12**, 617–637. (With Discussion).

Clements, M. P., and Hendry, D. F. (1993b). On the limitations of comparing mean squared forecast errors: A reply. *Journal of Forecasting*, **12**, 669–676.

Clements, M. P., and Hendry, D. F. (1995). Forecasting in cointegrated systems. *Journal of Applied Econometrics*, **10**, 127–146. Reprinted in Mills, T. C. (ed.) (1999), *Economic Forecasting*. The International Library of Critical Writings in Economics. Cheltenham: Edward Elgar.

Clements, M. P., and Hendry, D. F. (1996a). Forecasting in macro-economics. In Cox, D. R., Hinkley, D. V., and Barndorff-Nielsen, O. E. (eds.), *Time Series Models: In econometrics,*

finance and other fields, pp. 101–141. London: Chapman and Hall.

Clements, M. P., and Hendry, D. F. (1996b). Multi-step estimation for forecasting. *Oxford Bulletin of Economics and Statistics*, **58**, 657–684.

Clements, M. P., and Hendry, D. F. (1998a). Forecasting economic processes. *International Journal of Forecasting*, **14**, 111–131.

Clements, M. P., and Hendry, D. F. (1998b). *Forecasting Economic Time Series*. Cambridge: Cambridge University Press.

Clements, M. P., and Hendry, D. F. (1999). *Forecasting Non-stationary Economic Time Series*. Cambridge, Massachusetts: MIT Press.

Clements, M. P., and Hendry, D. F. (2001). Explaining the results of the M3 forecasting competition. *International Journal of Forecasting*, **17**, 550–554.

Clements, M. P., and Hendry, D. F. (eds.)(2002a). *A Companion to Economic Forecasting*. Oxford: Blackwells.

Clements, M. P., and Hendry, D. F. (2002b). Explaining forecast failure in macroeconomics. in Clements and Hendry (2002a), pp. 539–571.

Clements, M. P., and Hendry, D. F. (2003). Economic forecasting: Some lessons from recent research. *Economic Modelling*, **20**, 301–329.

Clements, M. P., and Hendry, D. F. (2005a). Guest editors' introduction: Information in economic forecasting. *Oxford Bulletin of Economics and Statistics*, **67**, 713–753.

Clements, M. P., and Hendry, D. F. (2005b). On intercept corrections worsening forecast performance. Unpublished paper, Nuffield College, Oxford University.

Clements, M. P., and Krolzig, H.-M. (1998). A comparison of the forecast performance of Markov-switching and threshold autoregressive models of US GNP. *Econometrics Journal*, **1**, C47–C75.

Clements, M. P., and Sensier, M. (2003). Asymmetric output-gap effects in Phillips curve and mark-up pricing models: Evidence for the US and the UK. *Scottish Journal of Political Economy*, **50**, 359–374.

Cochrane, J. H. (1994). Permanent and transitory components of GNP and stock prices. *Quarterly Journal of Economics*, **61**, 241–265.

Cockerell, L., and Russell, B. (1995). Australian wage and price inflation: 1971–1994. Research discussion paper No. 9509, Reserve Bank of Australia.

Coen, P. G., Gomme, E. D., and Kendall, M. G. (1969). Lagged relationships in economic forecasting. *Journal of the Royal Statistical Society A*, **132**, 133–163.

Coen, R. M., and Hickman, B. G. (2002). The productivity surge of the nineties and future growth. Discussion paper No. 01-26, Stanford Institute for Economic Policy Research.

Coyle, D. (2001). Making sense of published economic forecasts. in Hendry and Ericsson (2001), pp. 54–67.

Dahl, C. M., and Hylleberg, S. (2004). Flexible regression models and relative forecast performance. *International Journal of Forecasting*, **20**, 201–217.

Davidson, R., and MacKinnon, J. G. (1985). Heteroskedasticity-robust tests in regression directions. *Annales de l'INSEE*, **59/60**, 183–218.

De-Brouwer, G. (1998). Estimating output gaps. Discussion paper No. 9809, Reserve Bank of Australia.

De-Brouwer, G., and Ericsson, N. R. (1998). Modelling inflation in Australia. *Journal of Business and Economic Statistics*, **16**, 433–449.

De Gooijer, J. G., and Kumar, K. (1992). Some recent developments in non-linear time series modelling, testing and forecasting. *International Journal of Forecasting*, **8**, 135–156.

Dickey, D. A., and Fuller, W. A. (1981). Likelihood ratio statistics for autoregressive time series with a unit root. *Econometrica*, **49**, 1057–1072.

Dicks-Mireaux, L. A., and Dow, J. C. R. (1959). The determinants of wage inflation: United Kingdom, 1946-1956. *Journal of the Royal Statistical Society, A*, **122**, 145–184.

Diebold, F. X. (1988). Serial correlation and the combination of forecasts. *Journal of Business and Economic Statistics*, **6**, 105–111.

Diebold, F. X. (1998). *Elements of Forecasting*. Cincinnati: South-Western.

Diebold, F. X., Hahn, J. Y., 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**, 661–673.

Diebold, F. X., and Lopez, J. A. (1996). Forecast evaluation and combination. In Maddala, G. S., and Rao, C. R. (eds.), *Handbook of Statistics*, Vol. 14, pp. 241–268. Amsterdam: North-Holland.

Diebold, F. X., and Mariano, R. S. (1995). Comparing predictive accuracy. *Journal of Business and Economic Statistics*, **13**, 253–263.

Diebold, F. X., and Rudebusch, G. D. (1989). Scoring the leading indicators. *Journal of Business*, **62**, 369–391.

Diebold, F. X., and Rudebusch, G. D. (1991). Forecasting output with the composite leading index: An ex ante analysis. *Journal of the American Statistical Association*, **86**, 603–610.

Doan, T., Litterman, R., and Sims, C. A. (1984). Forecasting and conditional projection using realistic prior distributions. *Econometric Reviews*, **3**, 1–100.

Doornik, J. A. (1995). Econometric Computing. Ph.D Thesis, University of Oxford, Oxford.

Doornik, J. A. (1998). Approximations to the asymptotic distribution of cointegration tests. *Journal of Economic Surveys*, **12**, 573–593. Reprinted in M. McAleer and L. Oxley (1999). *Practical Issues in Cointegration Analysis*. Oxford: Blackwell Publishers.

Doornik, J. A. (2001). *Ox: An Object-Oriented Matrix Language*. London: Timberlake Consultants Press.

Doornik, J. A., and Hansen, H. (1994). A practical test for univariate and multivariate normality. Discussion paper, Nuffield College, Oxford University.

Doornik, J. A., and Hendry, D. F. (1996). *GiveWin: An Interface to Economic Modelling*. London: Timberlake Consultants Press.

Doornik, J. A., and Hendry, D. F. (2001a). *Empirical Econometric Modelling using PcGive: Volume I,* 3rd edn. London: Timberlake Consultants Press.

Doornik, J. A., and Hendry, D. F. (2001b). *Interactive Monte Carlo Experimentation in Econometrics using PcNaive*. London: Timberlake Consultants Press.

Doornik, J. A., and Hendry, D. F. (2001c). *Modelling Dynamic Systems using PcGive: Volume II,* 2nd edn. London: Timberlake Consultants Press.

Duesenberry, J. (1950). The mechanics of inflation. *Review of Economics and Statistics*, **32**, 144–149.

Dupasquier, C., Guay, A., and St-Amant, P. (1999). A survey of alternative methodologies for estimating potential output and the output gap. *Journal of Macroeconomics*, **21**, 577–595.

Durbin, J., and Watson, G. S. (1951). Testing for serial correlation in least squares regression II. *Biometrika*, **38**, 159–178.

Efroymson, M. A. (1960). Multiple regression analysis. In Ralston, A., and Wilf, H. S. (eds.), *Mathematical Methods for Digital Computers*, pp. 191–203. New York: John Wiley & Sons.

Emerson, R. A., and Hendry, D. F. (1996). An evaluation of forecasting using leading indicators. *Journal of Forecasting*, **15**, 271–291.

Engle, R. F. (1982). Autoregressive conditional heteroskedasticity with estimates of the variance of UK inflation. *Econometrica*, **50**, 987–1008.

Engle, R. F. (2002). New frontiers for ARCH models. *Journal of Applied Econometrics*, **17**, 425–446.

Engle, R. F., and Bollerslev, T. (1986). Modelling the persistence of conditional variances. *Econometric Reviews*, **5**, 1–50.

Engle, R. F., and Granger, C. W. J. (1987). Cointegration and error correction: Representation, estimation and testing. *Econometrica*, **55**, 251–276.

Engle, R. F., Hendry, D. F., and Richard, J.-F. (1983). Exogeneity. *Econometrica*, **51**, 277–304. Reprinted in Hendry (1993), pp. 330–357; Hendry (2000), pp. 330–357; and Ericsson and Irons (1994), pp. 39–70.

Engle, R. F., and Patton, A. (2001). What good is a volatility model?. *Quantitative Finance*, **1**, 237–245.

Engle, R. F., and White, H. (eds.)(1999). *Cointegration, Causality and Forecasting*. Oxford: Oxford University Press.

Engle, R. F., and Yoo, B. S. (1987). Forecasting and testing in co-integrated systems. *Journal of Econometrics*, **35**, 143–159.

Engle, R. F., and Yoo, B. S. (1991). Cointegrated economic time series: An overview with new results. In Engle, R. F., and Granger, C. W. J. (eds.), *Long-Run Economic Relationships*, pp. 237–266. Oxford: Oxford University Press.

Erdélyi, A. (1987). *Asymptotic Expansions*. New York: Dover.

Ericsson, N. R. (2002). Predictable uncertainty in economic forecasting. in Clements and Hendry (2002a), pp. 19–44.

Ericsson, N. R., Campos, J., and Tran, H.-A. (1990). PC-GIVE and David Hendry's econometric methodology. *Revista de Econometrica*, **10**, 7–117.

Ericsson, N. R., and Irons, J. S. (eds.)(1994). *Testing Exogeneity*. Oxford: Oxford University Press.

Espasa, A., Senra, E., and Albacete, R. (2002). Forecasting inflation in the European Monetary Union: A disaggregated approach by countries and by sectors. *European Journal of Finance*, **8(4)**, 402–421.

Evans, G. W., and Reichlin, L. (1994). Information, forecasts and the measurement of the business cycle. *Journal of Monetary Economics*, **33**, 233–254.

Fair, R. C. (1984). *Specification, Estimation and Analysis of Macroeconometric Models*. Massachusetts: Harvard University Press.

Faust, J., and Whiteman, C. H. (1995). Commentary on Grayham E. Mizon "Progressive modeling of macroeconomic times series: The LSE methodology". in Hoover (1995), pp. 171–180.

Faust, J., and Whiteman, C. H. (1997). General-to-specific procedures for fitting a data-admissible, theory-inspired, congruent, parsimonious, encompassing, weakly-exogenous, identified, structural model to the DGP: a translation and critique. *Carnegie-Rochester Conference Series on Economic Policy*, **47**, 121–161.

Fildes, R., and Ord, K. (2002). Forecasting competitions: Their role in improving forecasting practice and research. in Clements and Hendry (2002a), pp. 322–353.

Forni, M., Hallin, M., Lippi, M., and Reichlin, L. (2000). The generalized factor model: Identification and estimation. *The Review of Economics and Statistics*, **82**, 540–554.

Franses, P. H., and Teräsvirta, T. (2001). Nonlinear modeling of multivariate macroeconomic relations: Introduction. *Macroeconomic Dynamics*, **4**, 461–65.

Friedman, M. (1956). *Studies in the Quantity Theory of Money*. Chicago: University of Chicago Press.

Friedman, M. (1968). The role of monetary policy. *American Economic Review*, **58**, 1–17.

Friedman, M. (1970). A theoretical framework for monetary analysis. *Journal of Political Economy*, **78**, 193–238.

Friedman, M. (1971). A monetary theory of nominal income. *Journal of Political Economy*, **79**, 323–337.

Frisch, H. (1983). *Theories of Inflation*. Cambridge: Cambridge University Press.

Frisch, R. (1934). *Statistical Confluence Analysis by means of Complete Regression Systems*. Oslo: University Institute of Economics.

Frisch, R., and Waugh, F. V. (1933). Partial time regression as compared with individual trends. *Econometrica*, **1**, 387–401.

Furnival, G. M. (1971). All possible regressions with less computation. *Technometrics*, **13**, 403–

408.

Gallant, A. R. (1987). *Nonlinear Statistical Models*. New York: John Wiley & Sons.

Gardner, E. S. (1985). Exponential smoothing: the state of the art. *Journal of Forecasting*, **4**, 1–28.

Gatu, C., and Kontoghiorghes, E. J. (2006). Branch-and-bound algorithms for computing the best subset regression models. *Journal of Computational and Graphical Statistics*, **15**, 139–156.

Geriner, P. A., and Ord, J. K. (1991). Automatic forecasting using explanatory variables: A comparative study. *International Journal of Forecasting*, **7**, 127–140.

Gilbert, C. L. (1986). Professor Hendry's econometric methodology. *Oxford Bulletin of Economics and Statistics*, **48**, 283–307.

Gilbert, C. L. (1989). LSE and the British approach to time-series econometrics. *Oxford Economic Papers*, **41**, 108–128.

Giorno, C., Richardson, P., Roseveare, D., and van den Noord, P. (1995). Estimating potential output, output gaps and structural budget balances. Working paper No. 152, Organization for Economic Cooperation and Development.

Glosten, L. R., Jagannathan, R., and Runkle, D. E. (1993). On the relation between the expected value and the volatility of the nominal excess returns on stocks. *Journal of Finance*, **48**, 1779–1801.

Goldfeld, S. M., and Quandt, R. E. (1973). A Markov model for switching regressions. *Journal of Econometrics*, **1**, 3–16.

Gordon, R. J. (1997). The time-varying NAIRU and its implications for economic policy. *Journal of Economic Perspectives*, **11**, 11–32.

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. (1989). *Forecasting in Business and Economics,* 2nd edn. San Diego: Academic Press.

Granger, C. W. J. (ed.)(1990). *Modelling Economic Series*. Oxford: Clarendon Press.

Granger, C. W. J. (1993). Strategies for modelling nonlinear time series relationships. *Economic Record*, **60**, 233–238.

Granger, C. W. J. (2003). Evaluation of forecasts. in Hendry and Ericsson (2001), pp. 93–103.

Granger, C. W. J. (2005). Preface: Some thoughts on the future of forecasting. *Oxford Bulletin of Economics and Statistics*, **67**, 707–711.

Granger, C. W. J., and Lin, J.-L. (1994). Forecasting from nonlinear models in practice. *Journal of Forecasting*, **13**, 1–10.

Granger, C. W. J., and Newbold, P. (1973). Some comments on the evaluation of economic forecasts. *Applied Economics*, **5**, 35–47. Reprinted in Mills, T. C. (ed.) (1999), *Economic Forecasting*. The International Library of Critical Writings in Economics. Cheltenham: Edward Elgar.

Granger, C. W. J., and Newbold, P. (1986). *Forecasting Economic Time Series,* 2nd edn. New York: Academic Press.

Granger, C. W. J., and Pesaran, M. H. (2000a). A decision-based approach to forecast evaluation. In Chan, W. S., Li, W. K., and Tong, H. (eds.), *Statistics and Finance: An Interface*, pp. 261–278. London: Imperial College Press.

Granger, C. W. J., and Pesaran, M. H. (2000b). Economic and statistical measures of forecast accuracy. *Journal of Forecasting*, **19**, 537–560.

Granger, C. W. J., and Ramanathan, R. (1984). Improved methods of combining forecasting. *Journal of Forecasting*, **3**, 197–204.

Granger, C. W. J., and Sin, C.-Y. (2000). Modelling the absolute returns of different stock indices: Exploring the forecastability of an alternative measure of risk. *Journal of Forecasting*, **19**, 277–298.

Granger, C. W. J., and Teräsvirta, T. (1993). *Modelling Nonlinear Economic Relationships*.

Oxford: Oxford University Press.

Granger, C. W. J., and Teräsvirta, T. (1999). A simple nonlinear time series model with misleading linear properties. *Economics Letters*, **62**, 161–165.

Granger, C. W. J., and Uhlig, H. F. (1990). Reasonable extreme-bounds analysis. *Journal of Econometrics*, **44**, 159–170.

Green, P. J., and Silverman, B. W. (1994). *Nonparametric Regression and Generalized Linear Models. A Roughness Penalty Approach*. London: Chapman and Hall.

Griliches, Z., and Intriligator, M. (eds.)(1983). *Handbook of Econometrics*. Volume 1. Amsterdam: North Holland.

Griliches, Z., and Intriligator, M. (eds.)(1984). *Handbook of Econometrics*. Volume 2. Amsterdam: North Holland.

Haavelmo, T. (1944). The probability approach in econometrics. *Econometrica*, **12**, (supplement), 1–118.

Hackl, P., and Westlund, A. H. (eds.)(1991). *Economic Structural Change, Analysis and Forecasting*. Berlin: Springer-Verlag.

Hamilton, J. D. (1989). A new approach to economic analysis of nonstationary time series. *Econometrica*, **57**, 357–384.

Hamilton, J. D. (1994). *Time Series Analysis*. Princeton: Princeton University Press.

Hannan, E. J. (ed.)(1970). *Multiple Time Series*. New York: John Wiley & Sons.

Hannan, E. J., and Quinn, B. G. (1979). The determination of the order of an autoregression. *Journal of the Royal Statistical Society*, **B, 41**, 190–195.

Hansen, B. E. (1996). Methodology: Alchemy or science?. *Economic Journal*, **106**, 1398–1431.

Harbo, I., Johansen, S., Nielsen, B., and Rahbek, A. (1998). Asymptotic inference on cointegrating rank in partial systems. *Journal of Business and Economic Statistics*, **16(4)**, 388–399.

Harrison, R., Nikolov, K., Quinn, M., Ramsay, G., Scott, A., and Thomas, R. (2005). *The Bank of England Quarterly Model*. London: Bank of England.

Harvey, A. C. (1989). *Forecasting, Structural Time Series Models and the Kalman Filter*. Cambridge: Cambridge University Press.

Harvey, A. C. (1990). *The Econometric Analysis of Time Series,* 2nd edn. Hemel Hempstead: Philip Allan.

Harvey, A. C., and Jaeger, A. (1993). Detrending, stylized facts and the business cycle. *Journal of Applied Econometrics*, **8**, 231–247.

Harvey, A. C., and Todd, P. H. J. (1983). Forecasting economic time series with structural and Box-Jenkins models (with discussion). *Journal of Business and Economic Statistics*, **1**, 299–315.

Harvey, D., Leybourne, S., and Newbold, P. (1997). Testing the equality of prediction mean squared errors. *International Journal of Forecasting*, **13**, 281–291.

Harvey, D., Leybourne, S., and Newbold, P. (1998). Tests for forecast encompassing. *Journal of Business and Economic Statistics*, **16**, 254–259.

Harvey, D., Leybourne, S., and Newbold, P. (2000). Tests for multiple forecast encompassing. *Journal of Applied Econometrics*, **15**, 471–482.

Hatch, N. (2003). Modeling and forecasting at the Bank of England. in Hendry and Ericsson (2001), pp. 124–148.

Hendry, D. F. (1979). Predictive failure and econometric modelling in macro-economics: The transactions demand for money. In Ormerod, P. (ed.), *Economic Modelling*, pp. 217–242. London: Heinemann. Reprinted in Hendry (1993), pp. 246–269, and Hendry (2000), pp. 246–269.

Hendry, D. F. (1980). Econometrics: Alchemy or science?. *Economica*, **47**, 387–406. Reprinted in Hendry (1993), pp. 11–28, and Hendry (2000), pp. 11–28.

Hendry, D. F. (1986). The role of prediction in evaluating econometric models. *Proceedings of the Royal Society, A*, **407**, 25–33.

Hendry, D. F. (1993). *Econometrics: Alchemy or Science? Essays in Econometric Methodology*. Oxford: Blackwell Publishers.

Hendry, D. F. (1995). *Dynamic Econometrics*. Oxford: Oxford University Press.

Hendry, D. F. (1997a). The econometrics of macro-economic forecasting. *Economic Journal*, **107**, 1330–1357.

Hendry, D. F. (1997b). On congruent econometric relations: A comment. *Carnegie-Rochester Conference Series on Public Policy*, **47**, 163–190.

Hendry, D. F. (2000a). Does money determine UK inflation over the long run?. In Backhouse, R., and Salanti, A. (eds.), *Macroeconomics and the Real World. Volume 1*, pp. 85–114. Oxford: Oxford University Press.

Hendry, D. F. (2000b). *Econometrics: Alchemy or Science?* Oxford: Oxford University Press. New Edition.

Hendry, D. F. (2000c). Measuring the 'gap'. Mimeo, Nuffield College, Oxford University.

Hendry, D. F. (2001). Modelling UK inflation, 1875-1991. *Journal of Applied Econometrics*, **16**, 255–275.

Hendry, D. F. (2003). Denis Sargan and the origins of LSE econometric methodology. *Econometric Theory*, **19**, 456–480.

Hendry, D. F. (2005a). Equilibrium correction models. Mimeo, Nuffield College, Oxford University.

Hendry, D. F. (2005b). Robustifying forecasts from equilibrium-correction models. *Journal of Econometrics*. Special Issue in Honor of Clive Granger. Forthcoming.

Hendry, D. F., and Clements, M. P. (2000). Economic forecasting in the face of structural breaks. In Holly, S., and Weale, M. (eds.), *Econometric Modelling: Techniques and Applications*, pp. p3–37. Cambridge: Cambridge University Press.

Hendry, D. F., and Clements, M. P. (2004). Pooling of forecasts. *Econometrics Journal*, **7**, 1–31.

Hendry, D. F., and Clements, M. P. (2005). Forecasting using factor models. Mimeo, Department of Economics, Oxford University.

Hendry, D. F., and Ericsson, N. R. (eds.)(2001). *Understanding Economic Forecasts*. Cambridge, Massachusetts: MIT Press.

Hendry, D. F., and Hubrich, K. (2006). Forecasting aggregates by disaggregates. ECB Discussion Paper No. 589, European Central Bank.

Hendry, D. F., Johansen, S., and Santos, C. (2004). Selecting a regression saturated by indicators. Unpublished paper, Economics Department, University of Oxford.

Hendry, D. F., and Krolzig, H.-M. (1999). Improving on 'data mining reconsidered' by K.D. Hoover and S.J. Perez. *Econometrics Journal*, **2**, 41–58.

Hendry, D. F., and Krolzig, H.-M. (2001). *Automatic Econometric Model Selection*. London: Timberlake Consultants Press.

Hendry, D. F., and Krolzig, H.-M. (2003a). Model selection with more variables than observations. Unpublished paper, Economics Department, University of Oxford.

Hendry, D. F., and Krolzig, H.-M. (2003b). New developments in automatic general-to-specific modelling. in Stigum (2003), pp. 379–419.

Hendry, D. F., and Krolzig, H. M. (2004a). Automatic model selection: A new instrument for social science. *Electoral Studies*, **23**, 525–544.

Hendry, D. F., and Krolzig, H.-M. (2004b). Resolving three 'intractable' problems using a *Gets* approach. Unpublished paper, Economics Department, University of Oxford.

Hendry, D. F., and Krolzig, H.-M. (2004c). Sub-sample model selection procedures in general-to-specific modelling. In Becker, R., and Hurn, S. (eds.), *Contemporary Issues in Economics and Econometrics: Theory and Application*, pp. 53–75. Cheltenham: Edward Elgar.

Hendry, D. F., and Krolzig, H. M. (2004d). We ran one regression. *Oxford Bulletin of Economics and Statistics*, **66**, 799–810.

Hendry, D. F., and Krolzig, H.-M. (2005). The properties of automatic *Gets* modelling. *Economic*

Journal, **115**, C32–C61.

Hendry, D. F., and Mizon, G. E. (1990). Procrustean econometrics: or stretching and squeezing data. in Granger (1990), pp. 121–136.

Hendry, D. F., and Mizon, G. E. (1993). Evaluating dynamic econometric models by encompassing the VAR. In Phillips, P. C. B. (ed.), *Models, Methods and Applications of Econometrics*, pp. 272–300. Oxford: Basil Blackwell.

Hendry, D. F., and Mizon, G. E. (2000). Reformulating empirical macro-econometric modelling. *Oxford Review of Economic Policy*, **16**, 138–159.

Hendry, D. F., and Mizon, G. E. (2005). Forecasting in the presence of structural breaks and policy regime shifts. In Andrews, D. W., Powell, J. L., Ruud, P. A., and Stock, J. (eds.), *Identification and Inference in Econometric Models: Festschrift in Honor of T.J. Rothenberg*. Cambridge: Cambridge University Press. Forthcoming.

Hendry, D. F., and Morgan, M. S. (1989). A re-analysis of confluence analysis. *Oxford Economic Papers*, **41**, 35–52.

Hendry, D. F., and Morgan, M. S. (eds.)(1995). *The Foundations of Econometric Analysis*. Cambridge: Cambridge University Press.

Hendry, D. F., and Richard, J.-F. (1989). On the formulation of empirical models in dynamic econometrics. In Cornet, B., and Tulkens, H. (eds.), *Contributions to Operations Research and Economics: The Twentieth Anniversary of Core*. Cambridge, Massachusetts: MIT Press.

Hinich, M. J. (1982). Testing for gaussianity and linearity of a stationary time series. *Journal of Time Series Analysis*, **3**, 169–176.

Hocking, R. R., and Leslie, R. N. (1967). Selection of the best subset in regression analysis. *Technometrics*, **9**, 531–540.

Hodrick, R. J., and Prescott, E. C. (1997). Postwar US business cycles: An empirical investigation. *Journal of Money, Credit and Banking*, **29(1)**, 1–16. First published in 1980 as Carnegie Mellon University Discussion Paper No. 451.

Hoerl, A. E., and Kennard, R. W. (1970a). Ridge regression: Applications to nonorthogonal problems. *Technometrics*, **12**, 69–82.

Hoerl, A. E., and Kennard, R. W. (1970b). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, **12**, 55–67.

Hoeting, J. A., Madigan, D., Raftery, A. E., and Volinsky, C. T. (1999). Bayesian model averaging: A tutorial. *Statistical Science*, **14**, 382–417.

Hofmann, M., Gatu, C., and Kontoghiorghes, E. J. (2005). Efficient algorithms for computing the best-subset regression models for large scale problems. Presented at the IASC 3rd world conference on CSDA, Cyprus.

Holt, C. C. (1957). Forecasting seasonals and trends by exponentially weighted moving averages. ONR Research Memorandum 52, Carnegie Institute of Technology, Pittsburgh.

Hoover, K. D. (ed.)(1995). *Macroeconometrics: Developments, Tensions and Prospects*. Boston: Kluwer.

Hoover, K. D., and Perez, S. J. (1999). Data mining reconsidered: Encompassing and the general-to-specific approach to specification search. *Econometrics Journal*, **2**, 167–191.

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, **24**, 417–441.

Hubrich, K. (2005). Forecasting Euro area inflation: Does aggregating forecasts by HICP component improve forecast accuracy?. *International Journal of Forecasting*, **21**, 119–136.

James, W., and Stein, C. (1961). Estimation with quadratic loss. In Neyman, J. (ed.), *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, pp. 361–379. Berkeley: University of California Press.

Jerri, A. J. (1998). *The Gibbs Phenomenon in Fourier Analysis, Splines and Wavelet Approximations*. Dordrecht, Netherlands: Kluwer.

Johansen, S. (1988). Statistical analysis of cointegration vectors. *Journal of Economic Dynamics*

and Control, **12**, 231–254. Reprinted in Engle, R. F. and Granger, C. W. J. (eds.) (1991). *Long-Run Economic Relationships*, Oxford: Oxford University Press. pp.131-152.

Judge, G. G., and Bock, M. E. (1978). *The Statistical Implications of Pre-Test and Stein-Rule Estimators in Econometrics*. Amsterdam: North Holland.

Juselius, K. (1992). Domestic and foreign effects on prices in an open economy: The case of Denmark. *Journal of Policy Modeling*, **14**, 401–428.

Kaiser, H. F. (1960). The application of electronic computers to factor analysis. *Educational and Psychological Measurement*, **20**, 141–151.

Kalman, R. E. (1960). A new approach to linear filtering and prediction problems. *Transactions of the ASME–Journal of Basic Engineering, Series D*, **82**, 35–45.

Keenan, D. M. (1985). A Tukey non-additivity-type test for time series nonlinearity. *Biometrika*, **72**, 39–44.

Keynes, J. M. (1939). Professor Tinbergen's method. *Economic Journal*, **44**, 558–568.

Keynes, J. M. (1940). Comment. *Economic Journal*, **50**, 154–156.

Klein, L. R. (1971). *An Essay on the Theory of Economic Prediction*. Chicago: Markham Publishing Company.

Klein, L. R., Howrey, E. P., and MacCarthy, M. D. (1974). Notes on testing the predictive performance of econometric models. *International Economic Review*, **15**, 366–383.

Kohzadi, N., Boyd, M. S., Kaastra, I., Kermanshahi, B. S., and Scuse, D. (1995). Neural networks for forecasting: An introduction. *Canadian Journal of Agricultural Economics*, **43**, 463–474.

Koopman, S. J., Harvey, A. C., Doornik, J. A., and Shephard, N. (1995). *STAMP; Structural Time Series Analyser, Modeller and Predictor*. London: Timberlake Consultants Press.

Koopmans, T. C. (1947). Measurement without theory. *Review of Economics and Statistics*, **29**, 161–179.

Korenok, O., and Swanson, N. R. (2005). The incremental predictive information associated with using theoretical New Keynesian DSGE models vs. simple linear econometric models. *Oxford Bulletin of Economics and Statistics*, **67**, 905–930.

Körner, T. W. (1988). *Fourier Analysis*. Cambridge: Cambridge University Press.

Krolzig, H.-M., and Hendry, D. F. (2001). Computer automation of general-to-specific model selection procedures. *Journal of Economic Dynamics and Control*, **25**, 831–866.

Kuan, C.-M., and White, H. (1994). Artificial neural networks: An econometric perspective. *Econometric Reviews*, **13**, 1–91.

Kuttner, K. N. (1994). Estimating potential output as a latent variable. *Journal of Business and Economic Statistics*, **12**, 361–368.

Labhard, V., Kapetanios, G., and Price, S. (2005). Forecast combination and the Bank of England's suite of statistical forecasting models. Unpublished paper, Bank of England, London.

Lahiri, K., and Moore, G. H. (eds.)(1991). *Leading Economic Indicators: New Approaches and Forecasting Records*. Cambridge: Cambridge University Press.

Laidler, D. E. W., and Parkin, M. J. (1975). Inflation: A survey. *Economic Journal*, **85**, 741–809.

LaMotte, L. R., and Hocking, R. R. (1970). Computational efficiency in the selection of regression variables. *Technometrics*, **12**, 83–93.

Lamoureux, C. G., and Lastrapes, W. D. (1990). Persistence in variance, structural change, and the GARCH model. *Journal of Business and Economic Statistics*, **8**, 225–234.

Lanne, M., and Saikkonen, P. (2005). Non-linear GARCH models for highly persistent volatility. *Econometrics Journal*, **8**, 251–276.

Layard, R., Nickell, S. J., and Jackman, R. (1991). *Unemployment, Macroeconomic Performance and the Labour Market*. Oxford: Oxford University Press.

Leamer, E. E. (1978). *Specification Searches. Ad Hoc Inference with Non-Experimental Data*. New York: John Wiley & Sons.

Leamer, E. E. (1983a). Let's take the con out of econometrics. *American Economic Review*, **73**,

31–43. Reprinted in Granger (1990), pp. 29–49.

Leamer, E. E. (1983b). Model choice and specification analysis. in Griliches and Intriligator (1983), pp. 285–330.

Leamer, E. E. (1985). Sensitivity analyses would help. *American Economic Review*, **75**, 308–313.

Lee, T.-H., White, H., and Granger, C. W. J. (1993). Testing for neglected nonlinearity in time series models: A comparison of neural network methods and alternative tests. *Journal of Econometrics*, **56**, 269–290.

Li, W. K., Ling, S., and McAleer, M. (2002). Recent theoretical results for time series models with GARCH errors. *Journal of Economic Surveys*, **16**, 245–269.

Lin, J.-L., and Tsay, R. S. (1996). Co-integration constraint and forecasting. An empirical examination. *Journal of Applied Econometrics*, **11**, 519–538.

Lipsey, R. G. (1960). The relationship between unemployment and the rate of change of money wage rates in the UK, 1862–1957: A further analysis. *Economica*, **27(105)**, 1–32.

Ljung, G. M., and Box, G. E. P. (1978). On a measure of lack of fit in time series models. *Biometrika*, **66**, 67–72.

Lopez, J. (1999). Regulatory evaluation of Value-at-Risk models. *Journal of Risk*, **1**, 37–64.

Lorenz, H.-W. (1989). *Nonlinear Dynamic Economics and Chaotic Motion*. New York: Springer-Verlag.

Lovell, M. C. (1983). Data mining. *Review of Economics and Statistics*, **65**, 1–12.

Lucas, R. E. (1973). Some international evidence on output-inflation tradeoffs. *American Economic Review*, **63**, 326–334.

Lütkepohl, H. (1991). *Introduction to Multiple Time Series Analysis*. Berlin: Springer-Verlag.

Luukkonen, R., Saikkonen, P., and Teräsvirta, T. (1988). Testing linearity in univariate time series models. *Scandinavian Journal of Statistics*, **15**, 161–175.

Lynch, A. W., and Vital-Ahuja, T. (2003). Can subsample evidence alleviate the data-snooping problem? A comparison to the maximal R^2 cutoff test. Discussion paper, Stern Business School, New York University.

Maddala, G. S. (1977). *Econometrics*. New York: McGraw-Hill.

Makridakis, S., and Hibon, M. (2000). The M3-competition: Results, conclusions and implications. *International Journal of Forecasting*, **16**, 451–476.

Mallows, C. L. (1973). Some comments on C_p. *Technometrics*, **15**, 661–675.

Maravall, A. (1983). An application of nonlinear time series forecasting. *Journal of Business and Economic Statistics*, **1**, 66–74.

Marcellino, M. (2004). Forecasting EMU macroeconomic variables. *International Journal of Forecasting*, **20**, 359–372.

Marget, A. W. (1929). Morgenstern on the methodology of economic forecasting. *Journal of Political Economy*, **37**, 312–339.

Mariano, R. S. (2002). Testing forecast accuracy. in Clements and Hendry (2002a), pp. 284–298.

Marinucci, M. (2005). Retina winpack for real data. A quick guide for automatic model selection. http://personales.ya.com/max_mar/retina_winpack_v0b.pdf, Universidad Complutense de Madrid, Madrid.

McAleer, M., Pagan, A. R., and Volker, P. A. (1985). What will take the con out of econometrics?. *American Economic Review*, **95**, 293–301. Reprinted in Granger (1990), pp. 50–71.

McLeod, A. I., and Li, W. K. (1983). Diagnostic checking ARMA time series models using squared residual autocorrelations. *Journal of Time Series Analysis*, **4**, 269–273.

Metin, K. (1995). An integrated analysis of Turkish inflation. *Oxford Bulletin of Economics and Statistics*, **57**, 513–531.

Miller, A. (2002). *Subset Selection in Regression,* 2nd edn. New York: Chapman and Hall.

Mitchell, J., and Hall, S. G. (2005). Evaluating, comparing and combining density forecasts using the KLIC with and an application to the Bank of England and NIESR 'fan' charts of inflation. *Oxford Bulletin of Economics and Statistics*, **67**, 995–1033.

Mizon, G. E. (1994). The encompassing approach in econometrics. In Hendry, D. F., and Wallis, K. F. (eds.), *Econometrics and Quantitative Economics*, pp. 135–172. Oxford: Blackwells.

Mizon, G. E. (1995). Progressive modelling of macroeconomic time series: the LSE methodology. in Hoover (1995), pp. 107–170.

Mizon, G. E., and Richard, J.-F. (1986). The encompassing principle and its application to testing non-nested hypotheses. *Econometrica*, **54**, 657–678.

Morgan, J. A., and Tatar, J. F. (1972). Calculation of the residual sum of squares for all possible regressions. *Technometrics*, **14**, 317–325.

Morgan, W. A. (1939–1940). A test for the significance of the difference between the two variances in a sample from a normal bivariate population. *Biometrika*, **31**, 13–19.

Muellbauer, J. (1984). Aggregate production functions and productivity measurement: A new look. Discussion paper. No. 34, CEPR.

Muirhead, R. J. (1982). *Aspects of Multivariate Statistical Theory*. New York: John Wiley & Sons.

Naylor, T. H., Seaks, T. G., and Wichern, D. W. (1972). Box-Jenkins methods: An alternative to econometric models. *International Statistical Review*, **40**, 123–137.

Nelson, C. R. (1972). The prediction performance of the FRB-MIT-PENN model of the US economy. *American Economic Review*, **62**, 902–917. Reprinted in Mills, T. C. (ed.) (1999), *Economic Forecasting*. The International Library of Critical Writings in Economics. Cheltenham: Edward Elgar.

Nelson, D. B. (1991). Conditional heteroskedasticity in asset returns: A new approach. *Econometrica*, **59**, 347–370.

Nerlove, M. (1983). Expectations, plans and realizations in theory and practice. *Econometrica*, **51**, 1251–1279.

Newbold, P., and Harvey, D. I. (2002). Forecasting combination and encompassing. in Clements and Hendry (2002a), pp. 268–283.

Nicholls, D. F., and Quinn, B. G. (1982). *Random Coefficient Autoregressive Models: An Introduction*. Lecture Notes in Statistics, No. 11. New York: Springer-Verlag.

Nickell, S. J. (1990). Inflation and the UK labour market. *Oxford Review of Economic Policy*, **6**, 26–35.

Nielsen, B., and Rahbek, A. (2000). Similarity issues in cointegration models. *Oxford Bulletin of Economics and Statistics*, **62**, 5–22.

Nielsen, H. B., and Bowdler, C. (2003). Inflation adjustment in the open economy: An I(2) analysis of UK prices. Working paper, Nuffield College, University of Oxford.

Orphanides, A., and van Norden, S. (2002). The unreliability of output-gap estimates in real time. *Review of Economics and Statistics*, **84**, 569–583.

Pagan, A. R. (1987). Three econometric methodologies: a critical appraisal. *Journal of Economic Surveys*, **1**, 3–24.

Pagan, A. R., and Hall, A. D. (1983). Diagnostic tests as residual analysis. *Econometric Reviews*, **2**, 159–218.

Patterson, K. (2000). *An Introduction to Applied Econometrics. A Time Series Approach*. UK: Palgrave Macmillan.

Pedregal, D. J., and Young, P. C. (2002). Statistical approaches to modeling and forecasting time series. in Clements and Hendry (2002a), pp. 69–104.

Perez-Amaral, T., Gallo, G. M., and White, H. (2003). A flexible tool for model building: the relevant transformation of the inputs network approach (RETINA). *Oxford Bulletin of Economics and Statistics*, **65**, 821–838.

Perez-Amaral, T., Gallo, G. M., and White, H. (2005). A comparison of complementary automatic modelling methods: RETINA and PcGets. *Econometric Theory*, **21**, 262–277.

Pesaran, M. H., and Skouras, S. (2002). Decision-based methods for forecast evaluation. in Clements and Hendry (2002a), pp. 241–267.

Pesaran, M. H., Smith, R. P., and Yeo, J. S. (1985). Testing for structural stability and predictive failure: A review. *Manchester School*, **3**, 280–295.

Phelps, E. S. (1967). Phillips curves, expectations of inflation and optimal unemployment over time. *Economica*, **34**, 254–281.

Phillips, A. W. H. (1958). The relation between unemployment and the rate of change of money wage rates in the United Kingdom, 1861–1957. *Economica*, **25**, 283–299.

Phillips, A. W. H. (1962). Employment, inflation and growth. *Economica*, **29**, 1–16.

Phillips, P. C. B. (1988). Reflections on econometric methodology. *Economic Record*, **64**, 344–359.

Phillips, P. C. B. (1994). Bayes models and forecasts of Australian macroeconomic time series. In Hargreaves, C. (ed.), *Non-stationary Time Series Analyses and Cointegration*, pp. 53–86. Oxford: Oxford University Press.

Phillips, P. C. B. (1995). Automated forecasts of Asia-Pacific economic activity. *Asia-Pacific Economic Review*, **1**, 92–102.

Phillips, P. C. B. (1996). Econometric model determination. *Econometrica*, **64**, 763–812.

Phillips, P. C. B. (2003). Laws and limits of econometrics. *Economic Journal*, **113**, C26–C52.

Phillips, P. C. B., and Ploberger, W. (1994). Posterior odds testing for a unit root with data based model selection. *Econometric Theory*, **10**, 774–808.

Priestley, M. B. (1981). *Spectral Analysis and Time Series*. New York: Academic Press.

Proietti, T. (2002). Forecasting with structural time series models. in Clements and Hendry (2002a), pp. 105–132.

Proietti, T., and Harvey, A. C. (2000). A Beveridge-Nelson smoother. *Economic Letters*, **67**, 139–146.

Proietti, T., Musso, A., and Westermann, T. (2002). Estimating potential output and the output gap for the Euro area: a model-based production function approach. Working Paper No. ECO2002/09, European University Institute, Florence.

Puu, T. (1989). *Nonlinear Economic Dynamics*. New York: Springer-Verlag.

Quandt, R. E. (1983). Computational problems and methods. in Griliches and Intriligator (1983), pp. 699–746.

Quenouille, M. H. (1957). *The Analysis of Multiple Time Series*. Griffin's Statistical Monographs and Courses, edited by Kendall, M. G. London: Griffin.

Raftery, A. E., Painter, I., and Volinsky, C. T. (2005). BMA: An R package for Bayesian model averaging. *R News*, **5(2)**, 2–8.

Rahbek, A., and Mosconi, R. (1999). The role of stationary regressors in the cointegration test. *Econometrics Journal*, **2**, 76–91.

Ramsey, J. B. (1969). Tests for specification errors in classical linear least squares regression analysis. *Journal of the Royal Statistical Society B*, **31**, 350–371.

Rowlatt, P. A. (1988). Analysis of the recent path of UK inflation. *Oxford Bulletin of Economics and Statistics*, **50**, 335–360.

Rushton, S. (1951). On least squares fitting by orthogonal polynomials using the Choleski method. *Journal of the Royal Statistical Society B*, **13**, 92–99.

Saikkonen, P., and Luukkonen, R. (1988). Lagrange multiplier tests for testing nonlinearities in time series models. *Scandinavian Journal of Statistics*, **15**, 55–68.

Sargan, J. D. (1964). Wages and prices in the United Kingdom: A study in econometric methodology (with discussion). In Hart, P. E., Mills, G., and Whitaker, J. K. (eds.), *Econometric Analysis for National Economic Planning*, Vol. 16 of *Colston Papers*, pp. 25–63. London: Butterworth Co. Reprinted as pp. 275–314 in Hendry D. F. and Wallis K. F. (eds.) (1984). *Econometrics and Quantitative Economics*. Oxford: Basil Blackwell, and as pp. 124–169 in Sargan J. D. (1988), *Contributions to Econometrics*, Vol. 1, Cambridge: Cambridge University Press.

Sargan, J. D. (1980). A model of wage-price inflation. *Review of Economic Studies*, **47**, 979–1012. Reprinted as pp. 170–190 in Sargan J. D. (1988), *Contributions to Econometrics*, Vol. 1,

Cambridge: Cambridge University Press.

Sargan, J. D. (1982). On Monte Carlo estimates of moments that are infinite. In Basmann, R. L., and Rhodes, G. F. (eds.), *Advances in Econometrics: A Research Annual,*, Vol. 1, pp. 267–299. Greenwich, Connecticut: Jai Press Inc.

Sargan, J. D. (2001). Model building and data mining. *Econometric Reviews*, **20**, 159–170. First published (1973), Discussion Paper, London School of Economics.

Sargent, T. J., and Wallace, N. (1975). Rational expectations, the optimal monetary instrument, and the optimal money supply rule. *Journal of Political Economy*, **83**, 241–254.

Savin, N. E. (1984). Multiple hypothesis testing. in Griliches and Intriligator (1984), Ch. 14.

Schatzoff, M., Tsao, R., and Fienberg, S. (1968). Efficient calculation of all possible regressions. *Technometrics*, **10**, 769–779.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, **6**, 461–464.

Sensier, M., Artis, M., Osborn, D. R., and Birchenhall, C. (2004). Domestic and international influences on business cycle regimes in Europe. *International Journal of Forecasting*, **20**, 343–357.

Sims, C. A. (1980). Macroeconomics and reality. *Econometrica*, **48**, 1–48.

Staiger, D., Stock, J. H., and Watson, M. W. (1997). How precise are estimates of the natural rate of unemployment. In Romer, C., and Romer, D. (eds.), *Reducing Inflation: Motivation and Strategy*, Ch. 5. Chicago: Chicago University Press.

Stein, C. (1956). *Inadmissibility of the usual estimator for the mean of a multivariate normal distribution*. Berkeley: University of California Press.

Stigum, B. P. (ed.)(2003). *Econometrics and the Philosophy of Economics*. Princeton: Princeton University Press.

Stock, J. H., and Watson, M. W. (1989). New indexes of coincident and leading economic indicators. *NBER Macroeconomics Annual*, 351–409. London: MIT Press.

Stock, J. H., and Watson, M. W. (1993). A procedure for predicting recessions with leading indicators: Econometric issues and recent experience. In Stock, J. H., and Watson, M. W. (eds.), *Business Cycles, Indicators and Forecasting*, Ch. 2. Chicago: Chicago University Press.

Stock, J. H., and Watson, M. W. (1998). Diffusion indexes. Working paper No. 6702, NBER.

Stock, J. H., and Watson, M. W. (1999a). Business cycle fluctuations in US macroeconomic time series. In Taylor, J. B., and Woodford, M. (eds.), *Handbook of Macroeconomics*, Econometric Society Monograph 32. Amsterdam: North Holland.

Stock, J. H., and Watson, M. W. (1999b). A comparison of linear and nonlinear univariate models for forecasting macroeconomic time series. in Engle and White (1999), pp. 1–44.

Stock, J. H., and Watson, M. W. (1999c). Forecasting inflation. *Journal of Monetary Economics*, **44**, 293–335.

Stock, J. H., and Watson, M. W. (2002). Forecasting using principal components from a large number of predictors. *Journal of the American Statistical Association*, **97**, 1167–1179.

Subba Rao, T. (1985). On the theory of bilinear time series models. *Journal of the Royal Statistical Society, B*, **43**, 244–255.

Subba Rao, T., and Gabr, M. M. (1984). *Introduction to Bispectral Analysis and Bilinear Time Series Models*. New York: Springer-Verlag.

Swanson, N. R., and White, H. (1995). A model selection approach to assessing the information in the term structure using linear models and artificial neural networks. *Journal of Business and Economic Statistics*, **13**, 265–275.

Swanson, N. R., and White, H. (1997a). Forecasting economic time series using flexible versus fixed and linear versus nonlinear econometric models. *International Journal of Forecasting*, **13**, 439–461.

Swanson, N. R., and White, H. (1997b). A model selection approach to real-time macroeconomic forecasting using linear models and artificial neural networks. *Review of Economics and Statistics*, **79**, 540–550.

Tay, A. S., and Wallis, K. F. (2002). Density forecasting: A survey. in Clements and Hendry (2002a), pp. 45–68.

Teräsvirta, T. (1994). Specification, estimation and evaluation of smooth transition autoregressive models. *Journal of the American Statistical Association*, **89**, 208–218.

Teräsvirta, T. (1995). Modelling nonlinearity in U.S. gross national product 1889-1987. *Empirical Economics*, **20**, 577–597.

Teräsvirta, T. (1998). Modelling economic relationships with smooth transition regressions. In Ullah, A., and Giles, D. E. A. (eds.), *Handbook of Applied Economic Statistics*, pp. 507–552. New York: Marcel Dekker.

Teräsvirta, T., and Anderson, H. M. (1992). Characterizing nonlinearities in business cycles using smooth transition autoregressive models. *Journal of Applied Econometrics*, **7(S)**, S119–S136.

Teräsvirta, T., Tjostheim, D., and Granger, C. W. J. (1994). Aspects of modelling nonlinear time series. In Engle, R., and McFadden, D. (eds.), *Handbook of Econometrics*, pp. 2919–2960. Amsterdam: Elsevier.

Tiao, G. C. (1981). Modeling multiple time series with applications. *Journal of the American Statistical Association*, **76**, 802–816.

Tibshirani, R. (1996). Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society B*, **58**, 267–288.

Tinbergen, J. (ed.)(1940a). *Statistical Testing of Business Cycle Theories.* Vol. I: A Method and its applications to Investment Activity. Geneva: League of Nations.

Tinbergen, J. (ed.)(1940b). *Statistical Testing of Business Cycle Theories.* Vol. II: Business Cycles in the United States of America, 1919–1932. Geneva: League of Nations.

Tong, H. (1978). On a threshold model. In Chen, C. H. (ed.), *Pattern Recognition and Signal Processing*, pp. 101–141. Amsterdam: Sijhoff and Noordoff.

Tong, H. (1983). *Threshold Models in Non-Linear Time Series Analysis.* New York: Springer-Verlag.

Tong, H. (1990). *Non-Linear Time Series: A Dynamical System Approach.* Oxford: Oxford University Press.

Tsay, R. S. (1986). Non-linearity tests for time series. *Biometrika*, **73**, 461–466.

Tsay, R. S. (1987). Conditional heteroskedastic time series models. *Journal of the American Statistical Association*, **82**, 590–604.

Tsay, R. S. (1993). Comment: Adaptive forecasting. *Journal of Business and Economic Statistics*, **11**, 140–142.

Turner, D. (1995). Speed limit and asymmetric inflation effects from the output gap in the seven major economies. *OECD Economic Studies*, **24**, 57–88.

Turner, D. S. (1990). The role of judgement in macroeconomic forecasting. *Journal of Forecasting*, **9**, 315–345.

Wallis, K. F. (1999). Asymmetric density forecasts of inflation and the Bank of England's fan chart. *National Institute Economic Review*, **167**, 106–112.

Wallis, K. F. (2005a). An assessment of Bank of England and National Institute inflation forecast uncertainties. *National Institute Economic Review*, **198**, 64–71.

Wallis, K. F. (2005b). Combining density and interval forecasts: a modest proposal. *Oxford Bulletin of Economics and Statistics*, **67**, 983–994.

Wallis, K. F., Fisher, P. G., Longbottom, J., Turner, D. S., and Whitley, J. D. (1987). *Models of the UK Economy: A Fourth Review by the ESRC Macroeconomic Modelling Bureau.* Oxford: Oxford University Press.

Wallis, K. F., and Whitley, J. D. (1991). Sources of error in forecasts and expectations: UK economic models. *Journal of Forecasting*, **10**, 1984–1988.

Weiss, A. A. (1986). ARCH and bilinear time series models: combination and comparison. *Journal of Business and Economic Statistics*, **4**, 59–70.

West, K. D. (1996). Asymptotic inference about predictive ability. *Econometrica*, **64**, 1067–1084.

West, K. D. (2001). Tests for forecast encompassing when forecasts depend on estimated regression parameters. *Journal of Business and Economic Statistics*, **19**, 29–33.

West, K. D., and McCracken, M. W. (1998). Regression-based tests of predictive ability. *International Economic Review*, **39**, 817–840.

White, H. (1980). A heteroskedastic-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*, **48**, 817–838.

White, H. (1989). Some asymptotic results for learning in single hidden-layer feedforward network models. *Journal of the American Statistical Association*, **84**, 1003–1013.

Winters, P. R. (1960). Forecasting sales by exponentially weighted moving averages. *Management Science*, **6**, 324–342.

Wold, H. O. A. (1938). *A Study in the Analysis of Stationary Time Series.* Uppsala, Sweden: Almqvist & Wiksell.

Yancey, T. A., and Judge, G. G. (1976). A Monte Carlo comparison of traditional and Stein-rule estimators under squared error loss. *Journal of Econometrics*, **4**, 285–294.