

ESSAYS ON BAYESIAN MODEL AVERAGING  
USING ECONOMIC TIME SERIES

ISBN 978 90 5170 766 3

Cover design: Crasborn Graphic Designers bno, Valkenburg a.d. Geul

This book is no. **643** of the Tinbergen Institute Research Series, established through cooperation between Rozenberg Publishers and the Tinbergen Institute. A list of books which already appeared in the series can be found in the back.

# Essays on Bayesian Model Averaging using Economic Time Series

(Verhandelingen over Bayesiaanse modelmiddeling met economische tijdreeksen)

PROEFSCHRIFT

ter verkrijging van de graad van doctor  
aan de Erasmus Universiteit Rotterdam  
op gezag van de rector magnificus

Prof.dr. H.A.P. Pols

en volgens besluit van het College voor Promoties.

De openbare verdediging zal plaatsvinden op

donderdag 28 januari 2016 15.30 uur

door

RICHARD HUGO KLEIJN  
geboren te Sliedrecht.

## **Promotiecommissie**

Promotor: Prof.dr. H.K. van Dijk

Overige leden: Prof.dr. R. Paap  
Prof.dr. C.G. de Vries  
Dr. L.F. Hoogerheide

# Acknowledgements

Many persons have contributed in one way or another to the realization of this thesis during the many years that have passed since I started my PhD project. First of all, I would like to express my gratitude to my promotor Herman van Dijk. Without his support and optimism during all those years and especially in the period since last summer this thesis would not have existed. I am also very grateful to Frank Kleibergen and Richard Paap. I have learned a lot from you while working together on projects and in the many valuable discussions we have had on Bayesian econometrics. I also would like to thank my coauthors Lennart Hoogerheide, Francesco Ravazzolo, Nalan Baştürk and Marno Verbeek and the members of the doctoral committee.

During most of my time at the Tinbergen Institute I shared a room with Klaas Staal who has also been a great travel companion to several conferences and adventurous trips. I would also like to thank the other PhD students and staff at the Tinbergen and Econometric Institutes. Financial support by NWO (project 425-11-001 ‘Inference and decision-making in international economics and finance’) is gratefully acknowledged. I would like to thank my employer and my colleagues at PGGM for providing me the flexibility to work on finishing this thesis.

On a more personal note I would like to thank John en Bert for being such great friends, and to all my friends of the ‘zwemclub’ (and not only because they asked to be mentioned!): you are a great bunch of people. A final word of thanks must go to my beloved wife Marleen, to my parents, and to my brother Edwin for their love, their moral support, and for always being there for me. And above all, Soli Deo Gloria!

Richard Kleijn  
Sliedrecht, November 2015



# Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
1.1	Introduction . . . . .	2
1.1.1	Motivation . . . . .	2
1.1.2	Outline and Summary of Results . . . . .	2
1.2	Summary of chapters . . . . .	3
<b>2</b>	<b>Prior Ignorance, Likelihood Shape and Posterior Existence in a Cointegration Model</b>	<b>7</b>
2.1	Introduction . . . . .	8
2.2	A standard form of the cointegration model . . . . .	9
2.3	Linear normalization . . . . .	11
2.4	Existence of posterior moments under linear normalization with diffuse priors	17
2.4.1	Existence of the marginal posterior of $\beta_2 Y$ . . . . .	17
2.4.2	Existence of the marginal posterior of $\alpha Y$ . . . . .	18
2.5	Regularization through area restrictions using orthogonal normalization . .	23
2.5.1	Alternative specification and identification . . . . .	23
2.5.2	Prior choice and posterior moments . . . . .	25
2.6	Conclusions and remarks . . . . .	29
2.A	Results on matrices and determinants . . . . .	33
2.B	Matricvariate distributions . . . . .	35
2.B.1	Inverted Wishart distribution . . . . .	35
2.B.2	Matric-variate t distribution . . . . .	35
<b>3</b>	<b>Bayesian Analysis of the PPP Puzzle Using an Unobserved Components Model</b>	<b>37</b>
3.1	Introduction . . . . .	38
3.2	Data . . . . .	39
3.3	Model Specification . . . . .	40
3.4	Posterior simulator . . . . .	43
3.5	Posterior Results . . . . .	46
3.5.1	Posterior distribution of the model parameters . . . . .	46
3.5.2	Testing the sharp null hypothesis of stationarity . . . . .	46
3.5.3	Implied features of the model . . . . .	47
3.6	Final remarks . . . . .	51
3.A	Sampling the cycle parameters . . . . .	52

<b>4</b>	<b>Bayes Model Averaging of Cyclical Decompositions in Economic Time Series</b>	<b>55</b>
4.1	Introduction . . . . .	56
4.2	Cyclical Decomposition Procedure . . . . .	57
4.2.1	Specification . . . . .	58
4.2.2	Parameterization . . . . .	59
4.2.3	Prior . . . . .	60
4.2.4	Posterior Simulator . . . . .	61
4.2.5	Model Choice and Model Averaging . . . . .	62
4.2.6	Artificial Data Experiments . . . . .	63
4.3	US Industrial Production . . . . .	68
4.4	US Unemployment . . . . .	71
4.5	Summary and Conclusions . . . . .	74
4.A	Posterior sampler for a single cycle component . . . . .	76
4.B	Technical Details of the Posterior Simulator . . . . .	77
<b>5</b>	<b>Model Averaging using Time Varying Weights</b>	<b>79</b>
5.1	Introduction . . . . .	80
5.2	Forecast combination schemes . . . . .	81
5.2.1	Scheme 1: Bayesian Model Averaging (BMA) . . . . .	82
5.2.2	Combination schemes using estimated regression coefficients as model weights . . . . .	83
5.3	Financial application . . . . .	86
5.3.1	Evaluation . . . . .	87
5.3.2	Empirical Results . . . . .	89
5.4	US real GDP Growth . . . . .	92
5.5	Final remarks . . . . .	96
	<b>Nederlandse samenvatting (Summary in Dutch)</b>	<b>97</b>
	<b>Bibliography</b>	<b>101</b>



# Chapter 1

## Introduction

## 1.1 Introduction

### 1.1.1 Motivation

In many situations econometric inference may give different results for different models both in a Bayesian and a Frequentist approach. A common approach is that of model selection in which one single model is selected amongst a variety of possible models. All subsequent analysis such as forecasting is carried out using the selected model. The consequence is that information from the models that were not selected is ignored.

For this reason, in this thesis we not only consider model selection but also model averaging, which is an approach in which information from multiple models is combined. The motivating idea is that one single model will not be the best model under all circumstances. Combining these results, in particular for improved economic forecasting is very common nowadays. Recent economic forecasting and policy analysis shows a widespread interest in combining the information coming from different sources. One way of combining results is model averaging. Using model averaging information from all models considered is taken into account when forecasting. Model averaging is a key part of the Bayesian Approach and a very natural element of the Bayesian toolbox. Therefore in the Chapters 4 and 5 of this thesis essays are presented on BMA (Bayesian Model Averaging) applied to dynamic econometric models for economic time series, which address the issues mentioned above. In particular, we develop methods in which the weights of the models are allowed to evolve over time and investigate the contribution of this flexibility.

In order to obtain results several econometric tools and techniques have to be developed, which we do in Chapters 2 and 3.

### 1.1.2 Outline and Summary of Results

In the Bayesian approach a model weight is measured as a probability. Thus one deals with the evaluation of the prior and posterior probabilities of a model. We use here the assumption that the information in the scientific model dominates and thus we specify prior ignorance.

In Chapter 2 we investigate the implications of prior ignorance within the context of a cointegration model. In particular we explore the conditions under which posterior probabilities and higher moments exist. We propose an alternative parameter normalization which allows us to express ignorance as uniform priors on bounded sets.

In Chapter 3 we introduce a simulation-based Bayesian analysis of a particular unobserved components model which allows the mean of real exchange rates to be modelled as slowly time-varying. In the analysis we make use of Markov Chain Monte Carlo methods such as Gibbs sampling and Metropolis-Hastings, combined with the use of data augmentation which is a powerful method to deal with the unobserved components. In the Metropolis-Hastings step we use a carefully constructed candidate density. Subsequently, we use this model to construct the full implied posterior distributions of some key features of real exchange rates, specifically impulse responses. Bayes factors – which are closely linked to model probabilities – are calculated and

interpreted as a Bayesian stationarity test. We find that allowing the mean of the real exchange rates to be slowly time-varying supports the conclusions from Engel (2000). The decomposition between the permanent and transitory components helps to model these separate parts better which in turn is beneficial for modeling the real exchange rate series.

In Chapter 4 we introduce a novel class of time series models. This class of models provides a highly flexible specification with a variable number of stochastic cycles and it is closely related to the Fourier decomposition of a time series. We use the posterior simulator again to investigate the distribution of (highly nonlinear) functions of the model parameters. Next to that, we compute posterior model probabilities which on the one hand allow for testing for the number of cycle components and on the other hand these model probabilities enable a Bayesian Model Averaging approach over a wide range of specifications, involving varying numbers of cycle components and either including and excluding nonstationary components.

Chapter 5 extends the Bayesian Model Averaging approach with three different forecast combination schemes that allow for parameter uncertainty, model uncertainty and time varying model weights simultaneously. We consider two empirical applications, combining four models in the first case and combining six models in the second case. We investigate the performance of the averaging strategies in this context. We find that they give higher predictive quality than selecting the best model; secondly, we find that properly specified time varying model weights yield higher forecast accuracy and substantial economic gains compared with other averaging schemes.

This thesis contains some empirical applications to real exchange rates (Chapter 3), US industrial production and US unemployment time series (Chapter 4) and US stock market and US real GDP (Chapter 5). Although in some cases these are mainly illustrative for the applicability of the methods, we think these indicate future possibilities for applying of several of the results of this thesis.

## 1.2 Summary of chapters

### **Prior Ignorance, Likelihood Shape and Posterior Existence in a Cointegration Model**

Chapter 2 is based on Baştürk, Hoogerheide, Kleijn, and Van Dijk (2015, unpublished manuscript). In this chapter, shape and features of the posterior probability density function of the parameters of interest of a standard cointegration model under diffuse, improper priors are analyzed. These parameters are divided into two groups: cointegration parameters that refer to long run stable patterns in the time series and adjustment parameters that refer to error correction mechanisms that lead to a return to the long run patterns. Kleibergen and Van Dijk (1994) discuss possible pathological behavior of the posterior under a flat prior in this model. In the present chapter we extend this analysis and obtain, to the best of our knowledge, novel results:

(i) Using a flat prior in the standard multivariate regression model leads to a posterior of the regression coefficients that is a member of the well known multivariate

or matrix-variate  $t$  density. Using a flat prior in a standard cointegration model leads, however, to a posterior with a very nonstandard shape which refers to two typical features: the joint posterior may have a ridge in the surface and the tails may not go to zero sufficiently fast. This ridge implies that a marginal posterior may become unbounded in the interior of the parameter region. This feature starts to occur when the matrix of adjustment parameters tends to a reduced rank. That is, when the adjustment parameters tend to become weak in absolute value and even tend to zero. Then one is in the situation of near-unit root behavior of the economic time series. We show that, given the usual specification of a cointegration model, the marginal posterior of the adjustment parameters may become unbounded in the interior but that this asymptote is integrable. We also show that the tails of the posterior are too heavy and require extra information for regularity. These results are relevant for the specification of prior regularization information and also for the development of efficient computational procedures using Bayesian simulation-based methods.

(ii) We show that the existence of the posterior distribution of the cointegrating parameters is independent of the size of the model and of the number of cointegrating relations when linear normalization is used as the identification strategy. However, no first and higher order moments exist in the reduced rank case. We further show that the posterior distribution of the adjustment parameters is integrable on any finite area but it possesses too heavy tail behavior and is therefore not integrable on an infinite parameter space.

(iii) We develop a regularization approach in which the parameter space is transformed to bounded sets in a natural way. We build upon the concept of orthogonal normalization which is another identification strategy which is used in the cointegration literature and we extend this approach by introducing a modified parametrization and using a normalization that is closely associated with the singular value decomposition. This regularization then serves as the basis for the specification of prior ignorance.

Regularizing prior information based on weak and plausible restrictions on the range of the parameters of interest is introduced in order to obtain Bayesian inferential procedures for a cointegration model. Evaluating posterior probability distributions of parameters of a cointegration model is a crucial step in order to obtain inference on the structure of cointegration and on the level of the adjustment process. Equally important is being able to determine the weight given to a cointegration model. Once a model weight is obtained, Bayesian inference can proceed with model selection and model averaging, which is nowadays widely used in several fields of economics and finance, see, for instance, Chapter 5 of this thesis, Kleibergen and Paap (2002), Billio et al. (2013), Strachan and Van Dijk (2013) and Casarin et al. (2015).

## Bayesian Analysis of the PPP Hypothesis using an Unobserved Components Model

Chapter 3 is based on Kleijn and Van Dijk (2001). While the debate on whether long-run PPP holds or not will continue, we employ a pragmatic approach using an unobserved components model in which the importance of the mean reverting component relative to the non mean reverting component is allowed to vary. We provide some evidence for

short run PPP for the recent period of floating exchange rates between the end of the Bretton Woods system and the introduction of the euro. After correcting for the long-term variation, the half-lives drop to reasonable levels of about one year. It seems that replacing the fixed-mean assumption of PPP by a more flexible slowly varying mean is definitely helpful in this respect. Our Bayesian framework can fully quantify the uncertainty of the half-life estimates that Cheung and Lai (2000) report by providing the entire posterior density of the half-lives. We find the same kind of non-monotonicity in the impulse response function as they do. A shock is initially amplified before it starts to die out.

Of course, the approach taken in this chapter is pragmatic in nature but at least it suggests a direction in which an explanation for the PPP puzzle may be found. It still remains a challenging task to explain and model the mechanisms that determine the time-varying behavior of the real exchange rates. Non-traded goods might not be the only factors responsible for the time-variation of the equilibrium rate. A viable other possibility would be to use a monetary exchange rate model that includes interest rates and money demand and supply. This is an interesting topic of further research. Furthermore, in the PPP literature panel models have aided a lot to compensate for the limited time span of the recent period of floating exchange rates. Hence, an extension of our model to panels of countries is a possible topic for future research.

## Bayes Model Averaging of Cyclical Decompositions in Economic Time Series

Chapter 4 has been published in the *Journal of Applied Econometrics* as Kleijn and Van Dijk (2006). We have introduced a class of models, which consist of a stochastic trend, an *unknown* number of non-deterministic cyclical components and an irregular term. This specification leads to a very flexible functional form, which can capture a wide range of dynamics in the data. A MCMC-based Bayesian procedure is provided for sampling from the posterior distribution of the model parameters. Based on these samples, the posterior model probabilities can be calculated, indicating the appropriate number of cycles in the model. The parameter uncertainty and the uncertainty on the number of cycles can be fully taken into account in the subsequent inference. Prior specification is facilitated by a reparametrization and an efficient candidate generator for the model parameters of unobserved cycle components accelerates the MCMC-scheme.

Simulated data were used to illustrate model choice, model averaging, and that noise fitting is avoided. We have provided estimates of the cycles in US Industrial Production, extracted the evolution of the average growth rate over time and calculated forecast densities that incorporate all uncertainty originating from several sources. The asymmetry of the cycles in the US Unemployment rate data are adequately described by cyclical components at some selected higher order harmonic frequencies. Also, the underlying evolving mean of the unemployment rate and the amplitude of the cyclical variability were calculated.

A lot of further research in this area remains to be explored, including the implications of the flexibility of this approach for e.g. real time extraction of the structural component in a trend-cycle decomposition, the dating of business cycles and

other empirical applications. Possible model extensions to consider are multivariate analysis and explicit modelling of the synchronization mechanism between cycles.

## Forecast Accuracy and Economic Gains from Bayesian Model Averaging using Time Varying Weights

Chapter 5 has been published in *Journal of Forecasting* as Hoogerheide, Kleijn, Ravazzolo, van Dijk, and Verbeek (2010). This chapter contributes to the research on forecast combinations by investigating several Bayesian combination schemes. We propose four schemes that allow for parameter uncertainty, model uncertainty and time varying model weights simultaneously.

We provide two empirical illustrations. The results indicate that time varying model weight schemes outperform other averaging schemes in terms of predictive and economic gains. The first empirical example deals with forecasting the US stock market returns by combining individual forecasts from four competing models. The forecast combination schemes with time-varying model weights provide the highest economic gains. In particular, the forecast combination scheme with robust time varying weights provides much higher statistics and outperforms all the other models in terms of Sharpe ratio and realized utility value. The second empirical example refers to forecasting US economic growth over the business cycle, where we consider combinations of forecasts from six well-known time series models.

Our results suggest that model averaging may be very beneficial in business cycle analysis and forecasting, but as the number of individual models is relatively large and instability possibly high, weight estimates of the linear scheme may become inaccurate. Our results indicate that it is important that the averaging scheme is able to adapt quickly to the sharp decreases in GDP, particularly those in the 1980's and 1990's. We find that especially the scheme with robust time varying weights may early indicate recessions: before both the 1991 and 2001 crises its point forecast decreases substantially with approximately 0.5%.

The empirical applications have indicated, firstly, that averaging strategies can give higher predictive quality than selecting the best model; secondly, that properly specified time varying model weights yield higher forecast accuracy and substantial economic gains compared with other averaging schemes. The presented results lead to multiple directions for future research. As we already mentioned, interesting possibilities for further research are a rigorous analysis of the impact of some assumptions – both on theoretical aspects and practical applications – and an extensive study on the robustness of our findings.

## Chapter 2

# Prior Ignorance, Likelihood Shape and Posterior Existence in a Cointegration Model

*Chapter 2 is based on Baştürk, Hoogerheide, Kleijn, and Van Dijk (2015, unpublished manuscript).*

## 2.1 Introduction

‘Reporting the shape of the likelihood and its properties is an important task for a Bayesian econometrician’, was argued by Hildreth (1963). This viewpoint is based on the idea that scientific evidence should be reported in such a way that the information specified in the econometric model dominates with respect to other sources of information, see Baştürk et al. (2014b) for background and more details on this issue.

In this chapter, shape and features of the posterior probability density function of the parameters of interest of a standard cointegration model under diffuse, improper priors are analyzed. These parameters are divided into two groups: cointegration parameters that refer to long run stable patterns in the time series and adjustment parameters that refer to error correction mechanisms that lead to a return to the long run patterns. Given the indicated prior ignorance, the likelihood information strongly dominates. In a basic paper on this topic, Kleibergen and Van Dijk (1994) discuss possible pathological behavior of the posterior under a flat prior. In the present chapter we extend this analysis and obtain, to the best of our knowledge, novel results:

(i) Using a flat prior in the standard multivariate regression model leads to a posterior of the regression coefficients that is a member of the well known multivariate or matrix-variate  $t$  density. Using a flat prior in a standard cointegration model leads, however, to a posterior with a very nonstandard shape which refers to two typical features: the joint posterior may have a ridge in the surface and the tails may not go to zero sufficiently fast. This ridge implies that a marginal posterior may become unbounded in the interior of the parameter region. This feature starts to occur when the matrix of adjustment parameters tends to a reduced rank. That is, when the adjustment parameters tend to become weak in absolute value and even tend to zero. Then one is in the situation of near-unit root behavior of the economic time series. We show that, given the usual specification of a cointegration model, the marginal posterior of the adjustment parameters may become unbounded in the interior but that this asymptote is integrable. We also show that the tails of the posterior are too heavy and require extra information for regularity. These results are relevant for the specification of prior regularization information and also for the development of efficient computational procedures using Bayesian simulation-based methods. The latter has been an active area of research in Bayesian econometrics but it is a topic beyond the scope of this chapter. For an historical analysis of the development of this topic since the early nineteen-seventies we refer to Baştürk et al. (2014b).

(ii) Apart from analysis of the posterior shape, we show that the existence of the posterior distribution of the cointegrating parameters is independent of the size of the model and of the number of cointegrating relations when linear normalization is used as the identification strategy. However, no first and higher order moments exist in the reduced rank case. We further show that the posterior distribution of the adjustment parameters is integrable on any finite area but it possesses too heavy tail behavior and is therefore not integrable on an infinite parameter space.

(iii) We develop a regularization approach in which the parameter space is transformed to bounded sets in a natural way. We build upon the concept of orthogonal normalization which is another identification strategy which is used in the cointegration literature and we



extend this approach by introducing a modified parametrization and using a normalization that is closely associated with the singular value decomposition. This regularization then serves as the basis for the specification of prior ignorance.

Regularizing prior information based on weak and plausible restrictions on the range of the parameters of interest is introduced in order to obtain Bayesian inferential procedures for a cointegration model. Evaluating posterior probability distributions of parameters of a cointegration model is a crucial step in order to obtain inference on the structure of cointegration and on the level of the adjustment process. Equally important is being able to determine the weight given to a cointegration model. Once a model weight is obtained, Bayesian inference can proceed with model selection and model averaging, which is nowadays widely used in several fields of economics and finance, see, for instance, Chapter 5, Kleibergen and Paap (2002), Billio et al. (2013) Strachan and Van Dijk (2013) and Casarin et al. (2015).

As a remark, we note that the conditions to ensure the existence of moments that are obtained in this chapter are different from those of the IV model. The latter are given in Zellner, Ando, Baştürk, Hoogerheide, and Van Dijk (2014). In particular for the well-known horse model of two cointegrated time series the results are different from a basic IV model with one endogenous variable.

The above findings imply that adopting regularization priors is a useful strategy to establish the existence of the joint posterior distribution in cointegration models. Apart from the approach presented in this chapter, several other approaches on this topic have appeared in the literature. We present a brief survey in the final section but a detailed analysis of this literature is a topic outside the scope of this chapter.

The contents of this chapter are organized as follows. In Section 2.2 the standard cointegration model is presented. Conditional and marginal posterior densities of parameters of interest are derived under diffuse, improper priors in Section 2.3. Existence conditions for posterior moments are discussed in Section 2.4. In Section 2.5 the use of prior regularization information for the existence of moments is discussed. The final section contains our conclusions and some remarks on related results in the literature and on the meaning of the results for implied model features such as impulse response functions and multipliers. Technical details and background used in obtaining the results of this chapter are listed in the Appendix.

We emphasize that the present chapter is meant to give insight in shape and features of the posterior density and is not intended to yield the best functional form of such a density for structural analysis, forecasting and policy analysis using the Bayesian approach. More informative model and prior information is then needed.

## 2.2 A standard form of the cointegration model

A Vector AutoRegressive (VAR) model of lag order  $p$  is usually specified as

$$y_t = \mu + \tau t + \sum_{i=1}^p \Phi_i y_{t-i} + \varepsilon_t, \quad \varepsilon_t \sim NID(0, \Sigma), \quad \text{for } t = 1, \dots, T, \quad (2.1)$$

where  $y_t$  is  $k \times 1$  dimensional vector of observations on economic variables at time  $t$ ;  $\mu, \tau$  are  $k \times 1$  vectors of parameters of the deterministic components in the model;  $\Phi_i, i = 1, \dots, p$  are  $k \times k$  matrices of parameters belonging to the observations on the lagged endogenous variables; the disturbances  $\varepsilon_t$  for  $t = 1, \dots, T$  have independent Gaussian distributions with  $\Sigma$  as a positive definite symmetric (PDS) parameter matrix. Observations on  $y_0$  to  $y_{p-1}$  are given as initial values. For a general introduction to this class of models we refer to Johansen (1988).

The VAR model equation (2.1) can be cast into the Vector Error Correction Model (VECM) as follows:

$$\Delta y_t = \mu + \tau t + \Pi' y_{t-1} + \sum_{i=1}^{p-1} \Gamma_i \Delta y_{t-i} + \varepsilon_t, \quad (2.2)$$

where  $\Pi' = \sum_{i=1}^p \Phi_i - I_k$  and  $\Gamma_i = -\sum_{j=i+1}^p \Phi_j$  are  $k \times k$  matrices and  $I_k$  is the  $k \times k$  identity matrix.

The main motivation for the VECM specification is that it allows the user to focus on the issue whether the set of economic time series considered shows a return to long run stable relations or whether all series wander around like random walks. For more details and a broad exposition of this issue we refer to Johansen (1995).

We confine ourselves to the case of VARs of order 1, i.e.  $p = 1$ , without deterministic components. In this case (2.1), deleting the subindex 1, reduces to:

$$y_t = \Phi y_{t-1} + \varepsilon_t, \quad \varepsilon_t \sim NID(0, \Sigma), \text{ for } t = 1, \dots, T, \quad (2.3)$$

with the following error correction form:

$$\Delta y_t = \Pi' y_{t-1} + \varepsilon_t, \quad \varepsilon_t \sim NID(0, \Sigma), \text{ for } t = 1, \dots, T, \quad (2.4)$$

where  $\Pi' = \Phi - I_k$ . In matrix notation, this error correction model can be specified as:

$$\Delta Y = Y_{-1} \Pi + \mathcal{E}, \quad (2.5)$$

where  $\Delta Y$  is a  $T \times k$  matrix of observations  $\Delta y_1$  to  $\Delta y_T$  in its rows and similarly,  $Y_{-1}$  is a  $T \times k$  matrix of observations containing  $y_0$  to  $y_{T-1}$  in its rows. The  $T \times k$  random matrix  $\mathcal{E}$  has a matrix-variate distribution,  $\mathcal{E} \sim MN(0, \Sigma \otimes I_T)$ .

Using lag polynomial operators, the VAR model in Equation 2.3 can be written as:

$$A(L)y_t = \varepsilon_t, \quad (2.6)$$

where  $A(L) = I - \Phi L$ . Given the property that the class of lag polynomials is homomorphic with the class of real numbers, one can write  $A(z) = I - \Phi z$  with  $z \in \mathbb{R}$ . Note that in our case we have only a lag polynomial of order 1.

A necessary and sufficient condition for stationarity of  $y_t$  is that the characteristic polynomial  $A(z)$  in (2.6) has no unit roots, i.e.  $|A(1)| \neq 0$  under the further conditions that there are no seasonal unit roots, which correspond to complex unit roots with non-zero imaginary parts, and no explosive roots.

Since  $A(1) = I - \Phi = -\Pi'$ , stationarity of the model corresponds to  $\Pi$  having full rank. Then all series converge to a finite long run mean and have a bounded variance in the long run. When  $\Pi$  has rank 0, a  $k$ -dimensional random walk occurs. The long run mean is equal to the next period mean and long run variance tends to infinity. The more interesting case is where the process  $\{y_t\}$  has a so-called *cointegrating* rank  $r$ , that is, when  $\Pi$  has rank  $r$ . In this case one has  $r$  cointegrating or otherwise stated  $r$  stable relations between  $k$  economic variables and the matrix  $\Pi$  can be specified as the product of two  $k \times r$  matrices  $\alpha$  and  $\beta$  with full column rank and  $\Pi = \beta\alpha'$ .

The resulting model is called a *cointegrating* VECM, which in matrix notation takes the following form:

$$\Delta Y = Y_{-1}\beta\alpha' + \mathcal{E}, \quad (2.7)$$

The number of parameters in  $\alpha$  and  $\beta$  together may be larger than the number of free parameters in  $\Pi$  under a rank restriction. The following example shows that the same value of  $\Pi$  is obtained for infinitely many values of  $(\beta, \alpha)$  for  $k = 2$  and a reduced rank  $r = 1$ :

$$\Pi = \begin{pmatrix} 0.5 & 1 \\ 0.25 & 1 \end{pmatrix} = \begin{pmatrix} 1 \\ 0.5 \end{pmatrix} \begin{pmatrix} 0.5 & 1 \end{pmatrix} = \begin{pmatrix} c \\ 0.5c \end{pmatrix} \begin{pmatrix} 0.5/c & 1/c \end{pmatrix}, \quad \forall c \neq 0.$$

For the general case of  $k$  variables and  $r \leq k$  cointegrating relations, the same conclusion holds for any  $(r \times r)$  non-singular matrix  $R$ :

$$\Pi = \beta\alpha' = (\beta R)(\alpha R^{-1})',$$

with  $\text{rank}(\beta) = \text{rank}(\beta R)$  and  $\text{rank}(\alpha) = \text{rank}(\alpha R^{-1})$ . That is, the parameters  $\beta$  and  $\alpha$  are non-identified. A straightforward way of identifying the parameters is by using a linear normalization on  $\beta$  as restriction:

$$\beta = \begin{pmatrix} I_r \\ \beta_2 \end{pmatrix}, \quad (2.8)$$

where  $\beta_2$  is a  $(k-r) \times r$  matrix, see Kleibergen and Van Dijk (1994); Kleibergen and Paap (2002) among others. We consider as an alternative in Section 2.5 the case of orthogonal normalization.

Next to the general case, the 2-dimensional VAR,  $k = 2$  will be used for expository purposes. In partitioned form the VAR(2) model with one cointegrating relation can be written as:

$$\begin{pmatrix} \Delta y_{1,t} \\ \Delta y_{2,t} \end{pmatrix} = \begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix} \begin{pmatrix} 1 & \beta_2 \end{pmatrix} \begin{pmatrix} y_{1,t-1} \\ y_{2,t-1} \end{pmatrix} + \begin{pmatrix} \varepsilon_{1t} \\ \varepsilon_{2t} \end{pmatrix} \quad (2.9)$$

## 2.3 Posterior densities under linear normalization with diffuse priors

The likelihood function of the parameters of the model given the data is defined through its kernel as:

$$L(Y \mid \alpha, \beta_2, \Sigma) \propto |\Sigma|^{-T/2} \exp \left[ -\frac{1}{2} \text{tr} \left\{ \Sigma^{-1} (\Delta Y - Y_{-1}\beta\alpha')' (\Delta Y - Y_{-1}\beta\alpha') \right\} \right], \quad (2.10)$$

where the proportional sign  $\propto$  indicates that the right hand side of (2.10) equals the likelihood apart from an integrating constant.

We consider a diffuse class of priors defined on the space of  $(\alpha, \beta_2)$  and on the space of positive definite matrices  $\Sigma$ . Such a prior is defined as follows:

$$p(\beta_2, \alpha | Y) \propto |\Sigma|^{-h/2}, h > 1. \quad (2.11)$$

As a start we make use of the prior value  $h = k + 1$ , which gives an equivalence between the marginal posterior of  $(\alpha, \beta_2)$  and their, so-called, concentrated likelihood function. In Section 2.4 we discuss the effect of a more general choice of  $h$ .

The posterior density (apart from the integrating constant) under the normalization is evaluated by multiplying the likelihood in (2.10) and the prior in (2.11). This yields:

$$p(\alpha, \beta_2, \Sigma | Y) \propto |\Sigma|^{-(T+k+1)/2} \exp \left[ -\frac{1}{2} \text{tr} \{ \Sigma^{-1} (\Delta Y - Y_{-1} \beta \alpha')' (\Delta Y - Y_{-1} \beta \alpha') \} \right]. \quad (2.12)$$

**Marginal posterior of  $(\alpha, \beta_2)$ :** Using an inverse Wishart integration step on  $\Sigma$  in (2.12), the joint marginal density of  $\alpha$  and  $\beta_2$  is given by:

$$p(\alpha, \beta_2 | Y) \propto |(\Delta Y - Y_{-1} \beta \alpha')' (\Delta Y - Y_{-1} \beta \alpha')|^{-T/2}, \quad (2.13)$$

under the condition that  $(\Delta Y - Y_{-1} \beta \alpha')' (\Delta Y - Y_{-1} \beta \alpha')$  is a PDS matrix for all values of  $\alpha$  and  $\beta_2$  and  $T > k - 1$ . The first condition is satisfied when the data have full rank and the sample size requirement is usually not binding for time series in the field of econometrics. See Appendix 2.B.1 for more details on the definition of the inverted Wishart distribution and the inverse Wishart integration step.

The choice of  $h = k + 1$  implies that the information in the likelihood is the same from a Bayesian or a Frequentist perspective at this stage of the analysis. Possible sensitivity with respect to the choice of  $h$  is discussed in Section 2.4.

**Conditional posterior of  $(\alpha | \beta_2)$ :** The conditional posterior density of  $\alpha | \beta_2, Y$  is proportional to the joint posterior density  $p(\alpha, \beta_2 | Y)$  in (2.13)<sup>1</sup>:

$$p(\alpha | \beta_2, Y) \propto |(\Delta Y - Y_{-1} \beta \alpha')' (\Delta Y - Y_{-1} \beta \alpha')|^{-T/2}. \quad (2.14)$$

Completing the squares on  $\alpha$  in (2.14), using Appendix 2.A, yields:

$$p(\alpha | \beta_2, Y) \propto |(\Delta Y' M_{Y_{-1} \beta} \Delta Y + (\alpha - \hat{\alpha}) (\beta' Y_{-1}' Y_{-1} \beta) (\alpha - \hat{\alpha})')|^{-T/2}, \quad (2.15)$$

where  $\hat{\alpha}' = (\beta' Y_{-1}' Y_{-1} \beta)^{-1} \beta' Y_{-1}' \Delta Y$  and (2.15) holds under the conditions that  $\beta$  has full column rank, which is due to the normalization given in (2.8) and that  $\text{rank}(Y_{-1}) \geq r$ , hence the  $r \times r$  matrix  $(\beta' Y_{-1}' Y_{-1} \beta)$  has rank  $r$  and is invertible:

$$\text{rank}(\beta' Y_{-1}' Y_{-1} \beta) = \min(\text{rank}(Y_{-1}), r) = r. \quad (2.16)$$

---

<sup>1</sup>This relation holds since  $p(\alpha | \beta_2, Y) = \frac{p(\alpha, \beta_2 | Y)}{p(\beta_2 | Y)} \propto p(\alpha, \beta_2 | Y)$ .

From (2.15) and using the first definition of a matrixvariate  $t$  density given in Appendix 2.B.2, see also Dickey (1967), it follows that the conditional density of  $\alpha$  given  $\beta_2$  is proportional to a matrixvariate  $t$  density:

$$p(\alpha \mid \beta_2, Y) \propto p_{Mt}(\alpha \mid \hat{\alpha}, (\Delta Y' M_{Y_{-1}\beta} \Delta Y)^{-1}, (\beta' Y_{-1}' Y_{-1} \beta)^{-1}, T), \quad (2.17)$$

where the matrix  $\hat{\alpha}$  contains location parameter,  $\beta' Y_{-1}' Y_{-1} \beta$  and  $\Delta Y' M_{Y_{-1}\beta} \Delta Y$  are matrices that contain scale parameters and  $T > k + r - 1$  is a sample size requirement with  $T - k$  given as the degrees of freedom. For sample sizes that are usually given in econometrics, the latter condition is fulfilled. The matrixvariate  $t$  density property holds under the condition that  $\beta' Y_{-1}' Y_{-1} \beta$  and  $\Delta Y' M_{Y_{-1}\beta} \Delta Y$  are positive definite for all values of  $\beta_2$  which holds under linear normalization, see also below.

**Marginal posterior of  $\beta_2$**  From (2.14) and (2.17) we obtain:

$$\begin{aligned} p(\beta_2 \mid Y) &= \int p(\alpha, \beta_2 \mid Y) d\alpha \\ &\propto |\beta' Y_{-1}' Y_{-1} \beta|^{-k/2} |\Delta Y' M_{Y_{-1}\beta} \Delta Y|^{-(T-r)/2}. \end{aligned} \quad (2.18)$$

Using results listed in Appendix 2.A, the second factor in (2.18) can be written as:

$$|\Delta Y' M_{Y_{-1}\beta} \Delta Y| = \frac{|\beta' Y_{-1}' M_{\Delta Y} Y_{-1} \beta| |\Delta Y' \Delta Y|}{|\beta' Y_{-1}' Y_{-1} \beta|} \propto \frac{|\beta' Y_{-1}' M_{\Delta Y} Y_{-1} \beta|}{|\beta' Y_{-1}' Y_{-1} \beta|}. \quad (2.19)$$

Inserting (2.19) in (2.18), we obtain:

$$p(\beta_2 \mid Y) \propto \frac{|\beta' Y_{-1}' M_{\Delta Y} Y_{-1} \beta|^{-(T-r)/2}}{|\beta' Y_{-1}' Y_{-1} \beta|^{-(T-k-r)/2}}. \quad (2.20)$$

We next analyze the right hand side of (2.20) as function of  $\beta_2$  using the identification restrictions:  $\beta = (I \beta_2)'$ , hence  $Y_{-1}\beta = Y_{-1,1} + Y_{-1,2}\beta_2$  and thus the denominator becomes:

$$\beta' Y_{-1}' Y_{-1} \beta = (Y_{-1,1} + Y_{-1,2}\beta_2)' (Y_{-1,1} + Y_{-1,2}\beta_2). \quad (2.21)$$

Using similar results as presented in Appendix 2.A for completing the squares on  $\beta_2$  in the denominator of (2.20) yields:

$$\beta' Y_{-1}' Y_{-1} \beta = Y_{-1,1}' M_{Y_{-1,2}} Y_{-1,1} + (\beta_2 - \bar{\beta}_2)' Y_{-1,2}' Y_{-1,2} (\beta_2 - \bar{\beta}_2) \quad (2.22)$$

where

$$\bar{\beta}_2 = -(Y_{-1,2}' Y_{-1,2})^{-1} Y_{-1,2}' Y_{-1,1}, \quad (2.23)$$

Analogously, completing the squares on  $\beta_2$  in the numerator of (2.20) yields

$$\begin{aligned} \beta' Y_{-1}' M_{\Delta Y} Y_{-1} \beta &= Y_{-1,1}' M_{\Delta Y} Y_{-1,2} Y_{-1,1} \\ &\quad + (\beta_2 - \tilde{\beta}_2)' Y_{-1,2}' M_{\Delta Y} Y_{-1,2} (\beta_2 - \tilde{\beta}_2) \end{aligned} \quad (2.24)$$

where

$$\tilde{\beta}_2 = -(Y'_{-1,2} M_{\Delta Y} Y_{-1,2})^{-1} Y'_{-1,2} M_{\Delta Y} Y_{-1,1} \quad (2.25)$$

Using these two decompositions, the marginal posterior density of  $\beta_2$  in (2.20) is:

$$p(\beta_2 | Y) \propto \frac{\left| Y'_{-1,1} M_{\Delta Y} Y_{-1,2} Y_{-1,1} + (\beta_2 - \tilde{\beta}_2)' Y'_{-1,2} M_{\Delta Y} Y_{-1,2} (\beta_2 - \tilde{\beta}_2) \right|^{-(T-r)/2}}{\left| Y'_{-1,1} M_{Y_{-1,2}} Y_{-1,1} + (\beta_2 - \bar{\beta}_2)' Y'_{-1,2} Y_{-1,2} (\beta_2 - \bar{\beta}_2) \right|^{-(T-k-r)/2}}, \quad (2.26)$$

where  $\bar{\beta}_2$  and  $\tilde{\beta}_2$  are defined in (2.23) and (2.25), respectively.

**Proposition I** *Given the standard form of a cointegration model under linear normalization and using a diffuse class of priors, the marginal posterior of the cointegration parameters  $\beta_2$  is proportional to a matricvariate  $t$  density times a polynomial in  $\beta_2$ :*

$$p(\beta_2 | Y) \propto p_{Mt}(\beta_2 | \tilde{\beta}_2, Y'_{-1,2} M_{\Delta Y} Y_{-1,2}, Y'_{-1,1} M_{\Delta Y} Y_{-1,2} Y_{-1,1}, T-r) \times \left| Y'_{-1,1} M_{Y_{-1,2}} Y_{-1,1} + (\beta_2 - \bar{\beta}_2)' Y'_{-1,2} Y_{-1,2} (\beta_2 - \bar{\beta}_2) \right|^{(T-k-r)/2} \quad (2.27)$$

Conditions that guarantee that this is a proper density are discussed in Section 2.4.

**Conditional posterior of  $(\beta_2 | \alpha)$**  The conditional posterior density of  $\beta_2 | \alpha, Y$  is proportional to the joint posterior density  $p(\alpha, \beta_2 | Y)$  in (2.13).

This conditional is obtained in three steps. First, by completing the squares on  $\Pi = \beta \alpha'$ . Next, by completing the squares on  $\beta$  and thirdly by completing the squares on  $\beta_2$  and using the decomposition of the joint matricvariate  $t$  density of  $\beta$  into a conditional matricvariate  $t$  density of  $\beta_2$  and a marginal density of  $\beta_1$  evaluated at  $\beta_1 = I$ , see Appendix 2.B and Dickey (1967), Zellner (1971) or Bauwens et al. (1999) for background on the matricvariate  $t$  density.

The first step, completing the squares on  $\beta \alpha'$ , which is the restricted value of  $\Pi$ , proceeds as follows:

$$p(\alpha, \beta_2 | Y) \propto \left| (\Delta Y - Y_{-1} \beta \alpha')' (\Delta Y - Y_{-1} \beta \alpha') \right|^{-T/2}, \quad (2.28)$$

$$= \left| \Delta Y' M_{Y_{-1}} \Delta Y + (\beta \alpha' - \hat{\Pi})' Y'_{-1} Y_{-1} (\beta \alpha' - \hat{\Pi}) \right|^{-T/2}, \quad (2.29)$$

$$\propto \left| (Y'_{-1} Y_{-1})^{-1} + (\beta \alpha' - \hat{\Pi}) D^{-1} (\beta \alpha' - \hat{\Pi})' \right|^{-T/2} \quad (2.30)$$

where  $\hat{\Pi} = (Y'_{-1} Y_{-1})^{-1} Y'_{-1} \Delta Y$  and  $D = \Delta Y' M_{Y_{-1}} \Delta Y$ , which only depends on given data. In the last line we made use of the determinant equality in Appendix 2.A, equation (2.108).

The second step is completing the squares on  $\beta$  in (2.30). Here we use:

$$\hat{\beta} = \hat{\Pi} D^{-1} \alpha (\alpha' D^{-1} \alpha)^{-1}. \quad (2.31)$$

Hence (2.30) can be written as:

$$p(\alpha, \beta_2 | Y) \propto \left| (Y'_{-1} Y_{-1})^{-1} + (\beta - \hat{\beta}) \alpha' D^{-1} \alpha (\beta - \hat{\beta})' + \hat{\Pi} (D^{-1} - D^{-1} \alpha (\alpha' D^{-1} \alpha)^{-1} \alpha' D^{-1}) \hat{\Pi}' \right|^{-T/2}, \quad (2.32)$$

where the term  $\hat{\Pi} (D^{-1} - D^{-1} \alpha (\alpha' D^{-1} \alpha)^{-1} \alpha' D^{-1}) \hat{\Pi}'$  results from the difference between  $(\beta \alpha' - \hat{\Pi}) D^{-1} (\beta \alpha' - \hat{\Pi})'$  from (2.30) and  $(\beta - \hat{\beta}) \alpha' D^{-1} \alpha (\beta - \hat{\beta})'$  from (2.32).

Using Appendix 2.A, equation (2.93), (2.32) can be simplified as follows:

$$\begin{aligned} p(\alpha, \beta_2 | Y) &\propto \left| (\beta - \hat{\beta}) \alpha' D^{-1} \alpha (\beta - \hat{\beta})' + (Y'_{-1} Y_{-1} - Y'_{-1} \Delta Y \alpha_{\perp} (\alpha'_{\perp} \Delta Y' \Delta Y \alpha_{\perp})^{-1} \alpha'_{\perp} \Delta Y' Y_{-1})^{-1} \right|^{-T/2} \\ &= \left| (Y'_{-1} M_{\Delta Y \alpha_{\perp}} Y_{-1})^{-1} + (\beta - \hat{\beta}) \alpha' D^{-1} \alpha (\beta - \hat{\beta})' \right|^{-T/2}, \end{aligned} \quad (2.33)$$

$$(2.34)$$

where  $r \times r$  values of  $\beta$  are fixed due to the normalization restriction and the orthogonal complement  $\alpha_{\perp}$  satisfies  $\alpha' \alpha_{\perp} = 0$  and  $\alpha'_{\perp} \alpha_{\perp} = I$ . Equation (2.34) takes the form of a matrixvariate  $t$  density for the unrestricted parameter matrix  $\beta = (\beta'_1, \beta'_2)'$ , when the linear normalization is not used. Note that the matrices  $Y'_{-1} M_{\Delta Y \alpha_{\perp}} Y_{-1}$  and  $\alpha' D^{-1} \alpha$  are required to be positive definite for all values of the random variable matrix  $\alpha$ . If one or more columns of  $\alpha$  go to zero, then the matrix  $\alpha' D^{-1} \alpha$  becomes singular. We can ignore that here since this event has probability measure zero within the space of  $\alpha$ . When columns are very close but not equal to zero, then the matrix is nearly singular. We investigate the limiting behavior when the columns become arbitrarily close to zero in the next section. This is a situation of near-unit roots which is an empirical relevant issue that received a lot of attention in the literature. We comment on this also in the conclusions.

We next make use of the decomposition of a matrixvariate  $t$  density into a conditional and marginal one, as mentioned before:

$$\begin{aligned} p(\alpha, \beta | Y) &\propto \left| (Y'_{-1} M_{\Delta Y \alpha_{\perp}} Y_{-1})^{-1} + (\beta - \hat{\beta}) \alpha' D^{-1} \alpha (\beta - \hat{\beta})' \right|^{-T/2} \\ &= \left| P^{-1} + (\beta - \hat{\beta}) Q^{-1} (\beta - \hat{\beta})' \right|^{-T/2} \end{aligned} \quad (2.35)$$

$$\stackrel{(2.107)}{=} \frac{|P|^{(T-r)/2} |Q|^{k/2}}{c(k, r, T)} p_{Mt}(\beta | \hat{\beta}, P, Q, T) \quad (2.36)$$

$$\begin{aligned} &\stackrel{(2.112, 2.113)}{=} \frac{|P|^{(T-r)/2} |Q|^{k/2}}{c(k, r, T)} p_{Mt}(\beta_2 | \hat{\beta}_{2|1}, P_{22}, Q_{2|1}, T) \\ &\quad \times p_{Mt}(\beta_1 | \hat{\beta}_1, P_{11.2}, Q, T - k + r) \end{aligned} \quad (2.37)$$

$$\begin{aligned}
P &= Y_{-1}' M_{\Delta Y \alpha_{\perp}} Y_{-1} \\
Q &= (\alpha' D^{-1} \alpha)^{-1} \\
c(k, r, T) &= \frac{1}{\pi^{kr/2}} \frac{\prod_{i=1}^r \Gamma[(T-i+1)/2]}{\prod_{i=1}^r \Gamma[(T-k-i+1)/2]} \\
P_{11.2} &= P_{11} - P_{12} P_{22}^{-1} P_{21} \\
Q_{2|1} &= Q + (\beta_1 - \hat{\beta}_1)' P_{11.2} (\beta_1 - \hat{\beta}_1) \\
\hat{\beta}_{2|1} &= \hat{\beta}_2 - P_{22}^{-1} P_{21} (\beta_1 - \hat{\beta}_1)
\end{aligned}$$

Since  $p(\beta_2 | \alpha, Y)$  is proportional to  $p(\alpha, \beta_2 | Y)$  and  $p_{Mt}(\beta_2 | \hat{\beta}_{2|1}, P_{22}, Q_{2|1}, T)$  is the only factor that depends on  $\beta_2$  it follows that

$$p(\beta_2 | \alpha, Y) \propto p_{Mt}(\beta_2 | \hat{\beta}_{2|1}, P_{22}, Q_{2|1}, T) \quad (2.38)$$

$$\propto |Q_{2|1}|^{(T-k+r)/2} |P_{22}|^{r/2} \left| Q_{2|1} + (\beta_2 - \hat{\beta}_{2|1})' P_{22} (\beta_2 - \hat{\beta}_{2|1}) \right|^{-T/2} \quad (2.39)$$

Now we can integrate out  $\beta_2$  and evaluate the marginal density of  $\beta_1$  at the normalization restriction  $\beta_1 = I_r$ :

$$\begin{aligned}
p(\alpha | Y) &= \int p(\alpha, \beta | Y) d\beta_2 \\
&\propto |P|^{(T-r)/2} |Q|^{k/2} p_{Mt}(\beta_1 | \hat{\beta}_1, P_{11.2}, Q, T - k + r) \quad (2.40)
\end{aligned}$$

$$\begin{aligned}
&= |P|^{(T-r)/2} |Q|^{k/2} |P_{11.2}|^{r/2} |Q|^{(T-k+r-r)/2} \\
&\quad \times \left| Q + (\beta_1 - \hat{\beta}_1)' P_{11.2} (\beta_1 - \hat{\beta}_1) \right|^{-(T-k+r)/2} \quad (2.41)
\end{aligned}$$

$$\begin{aligned}
&\stackrel{\beta_1 = I_r}{=} |P|^{(T-r)/2} |Q|^{T/2} |P_{11} - P_{12} P_{22}^{-1} P_{21}|^{r/2} \\
&\quad \times \left| Q + (I_r - \hat{\beta}_1)' (P_{11} - P_{12} P_{22}^{-1} P_{21}) (I_r - \hat{\beta}_1) \right|^{-(T-k+r)/2} \quad (2.42)
\end{aligned}$$

$$\begin{aligned}
&= |P|^{T/2} |P_{22}|^{-r/2} |Q|^{T/2} \\
&\quad \times \left| Q + (I_r - \hat{\beta}_1)' (P_{11} - P_{12} P_{22}^{-1} P_{21}) (I_r - \hat{\beta}_1) \right|^{-(T-k+r)/2} \quad (2.43)
\end{aligned}$$

where we have used  $|P_{11.2}| = |P_{11} - P_{12} P_{22}^{-1} P_{21}| = |P|/|P_{22}|$ .

**Proposition II** *Given the standard form of a cointegration model under linear normalization and using a diffuse class of priors, the marginal posterior density of the adjustment parameters  $\alpha$  is a rational function of  $\alpha$ , given as:*

$$p(\alpha | Y) \propto |P|^{T/2} |P_{22}|^{-r/2} |Q|^{T/2} \left| Q + (I_r - \hat{\beta}_1)' P_{11.2} (I_r - \hat{\beta}_1) \right|^{-(T-k+r)/2}, \quad (2.44)$$

*and this density is not proportional to a known form of densities.*

Whether this is a proper density is analyzed in the next section.

Remark: The posterior densities given in (2.27) and (2.44) have the same functional form as the ones given in Kleibergen and Van Dijk (1994), equations (28) and (29). We



note here again the restriction that  $Q = (\alpha' D^{-1} \alpha)^{-1}$  is pds for all values of  $\alpha$  in the real space and that the event of a reduced rank of the matrix  $\alpha$ , in particular the event  $\alpha = 0$ , is an event with probability measure zero. It is, however, important to note what happens with the shape of the densities when  $\alpha$  tends to and is close to zero.

## 2.4 Existence of posterior moments under linear normalization with diffuse priors

### 2.4.1 Existence of the marginal posterior of $\beta_2 | Y$

We first rewrite (2.20) as follows

$$p(\beta_2 | Y) \propto |\beta' Y'_{-1} Y_{-1} \beta|^{-k/2} \left( \frac{|\beta' Y'_{-1} M_{\Delta Y} Y_{-1} \beta|}{|\beta' Y'_{-1} Y_{-1} \beta|} \right)^{-(T-r)/2} \quad (2.45)$$

where the second factor is a matrix generalization of a Rayleigh quotient. Similarly to the vector case, also in this matrix case the quotient is bounded from above and below by functions of the given data which we can show by defining

$$B = (Y'_{-1} Y_{-1})^{1/2} \beta (\beta' Y'_{-1} Y_{-1} \beta)^{-1/2} \quad (2.46)$$

$$W = (Y'_{-1} Y_{-1})^{-1/2} Y'_{-1} M_{\Delta Y} Y_{-1} (Y'_{-1} Y_{-1})^{-1/2} \quad (2.47)$$

and rewriting the quotient as:

$$\begin{aligned} \frac{|\beta' Y'_{-1} M_{\Delta Y} Y_{-1} \beta|}{|\beta' Y'_{-1} Y_{-1} \beta|} &= |\beta' Y'_{-1} Y_{-1} \beta|^{-1/2} |\beta' Y'_{-1} M_{\Delta Y} Y_{-1} \beta| |\beta' Y'_{-1} Y_{-1} \beta|^{-1/2} \\ &= |(\beta' Y'_{-1} Y_{-1} \beta)^{-1/2} \beta' Y'_{-1} M_{\Delta Y} Y_{-1} \beta (\beta' Y'_{-1} Y_{-1} \beta)^{-1/2}| \\ &= |B' W B|. \end{aligned} \quad (2.48)$$

Since it holds that  $B' B = I_r$  we can apply Theorem 11.15 from Magnus and Neudecker (1995) which states that  $|B' W B|$  is bounded having the product of the  $r$  smallest eigenvalues of the matrix  $W$  as its lower bound and the product of the  $r$  largest eigenvalues of  $W$  as its upper bound.

Therefore, integrability of the function (2.45) depends on the integrability of the factor  $|\beta' Y'_{-1} Y_{-1} \beta|^{-k/2}$ . Using (2.22) we rewrite the integral as

$$\begin{aligned} \int |\beta' Y'_{-1} Y_{-1} \beta|^{-k/2} d\beta_2 &= \\ \int |Y'_{-1,1} M_{Y_{-1,2}} Y_{-1,1} + (\beta_2 - \bar{\beta}_2)' Y'_{-1,2} Y_{-1,2} (\beta_2 - \bar{\beta}_2)|^{-k/2} d\beta_2. \end{aligned} \quad (2.49)$$

The integrand is proportional to a matrixvariate  $t$  density with  $r$  degrees of freedom which exists under the condition that  $k > (k - r) + r - 1 = k - 1$ .

**Proposition III** *Given the standard form of a cointegration model under linear normalization and using a diffuse class of priors, the marginal posterior distribution of the cointegration parameters  $\beta_2$ , with density (2.49), exists independent of the cointegrating rank  $r$ , but no first or higher order moments exist.*

It is noteworthy that this result is also independent of the difference  $k - r$ . We come back to this point in the next subsection. This result extends the analysis and results of Kleibergen and Van Dijk (1994).

Remark: Here we mention that the choice of the prior parameter  $h$  does not play a role in the existence condition for the function (2.45).

**Marginal posterior of  $\beta_2$  for  $k = 2, r = 1$**  For the special case  $k = 2, r = 1$ , positive definiteness of the left hand side of (2.49) is trivial if for convenience the data matrices are scaled and rotated such that  $Y'_{-1}Y_{-1} = I_k$ :

$$\int |\beta' Y'_{-1} Y_{-1} \beta|^{-k/2} d\beta_2 = \int (1 + \beta_2^2)^{-1} d\beta_2 \quad (2.50)$$

The integrand is proportional to a Cauchy density. Hence, the integral is finite and the marginal posterior of  $\beta_2$  exists but no finite first or higher order moments.

## 2.4.2 Existence of the marginal posterior of $\alpha|Y$

By (2.44) the marginal posterior density of  $\alpha | Y$  is as follows:

$$p(\alpha | Y) \propto |P|^{T/2} |P_{22}|^{-r/2} |Q|^{T/2} \left| Q + (I_r - \hat{\beta}_1)' P_{11.2} (I_r - \hat{\beta}_1) \right|^{-(T-k+r)/2}. \quad (2.51)$$

For the existence conditions of the distribution with this density, we first show that the first two factors in the right hand side of (2.51) are bounded. Consider:

$$|P_{22}|^{-r/2} |P|^{T/2} = |(Y'_{-1} M_{\Delta Y \alpha_{\perp}} Y_{-1})_{\{l, k-r\}}|^{-r/2} |Y'_{-1} M_{\Delta Y \alpha_{\perp}} Y_{-1}|^{T/2} \quad (2.52)$$

where  $A_{\{l, b\}}$  denotes the  $b \times b$  lower-right minor of matrix  $A$ . From Theorem 11.16 in Magnus and Neudecker (1995) we have that  $|(Y'_{-1} M_{\Delta Y \alpha_{\perp}} Y_{-1})_{\{l, k-r\}}|$  has its lower bound equal to the product of the  $k - r$  smallest eigenvalues of  $Y'_{-1} M_{\Delta Y \alpha_{\perp}} Y_{-1}$  and its upper bound is equal to the product of the  $k - r$  largest eigenvalues. Note that the matrix  $Y'_{-1} M_{\Delta Y \alpha_{\perp}} Y_{-1}$  is positive definite in the typical set up of econometrics, *e.g.*  $\text{rank}(M_{\Delta Y \alpha_{\perp}}) = T - \text{rank}(\Delta Y \alpha_{\perp}) \geq T - r$  and hence these products of eigenvalues are bounded.

Using (2.101) we have

$$|Y'_{-1} M_{\Delta Y \alpha_{\perp}} Y_{-1}| = \frac{|\alpha'_{\perp} \Delta Y' M_{Y_{-1}} \Delta Y \alpha_{\perp}|}{|\alpha'_{\perp} \Delta Y' \Delta Y \alpha_{\perp}|} |Y'_{-1} Y_{-1}| \quad (2.53)$$

of which the last factor is constant given the data and the first factor is bounded by products of the  $r$  smallest and largest eigenvalues of

$$(\Delta Y' \Delta Y)^{-1/2} \Delta Y' M_{Y_{-1}} \Delta Y (\Delta Y' \Delta Y)^{-1/2} \quad (2.54)$$

by similar arguments as in subsection 2.4.1.

Hence the density in (2.51) integrates to a finite value if the product of the last two factors  $|Q|^{T/2} \left| Q + (I_r - \hat{\beta}_1)' P_{11.2} (I_r - \hat{\beta}_1) \right|^{-(T-k+r)/2}$  has a finite integral. Note again that  $Q$  here is a function of  $\alpha$  with  $Q = (\alpha' D^{-1} \alpha)^{-1}$ .

Since  $(I_r - \hat{\beta}_1)' P_{11.2} (I_r - \hat{\beta}_1)$  is positive semidefinite and therefore

$$|Q| \leq \left| Q + (I_r - \hat{\beta}_1)' P_{11.2} (I_r - \hat{\beta}_1) \right| \quad (2.55)$$

we have that

$$\begin{aligned} |Q|^{T/2} \left| Q + (I_r - \hat{\beta}_1)' P_{11.2} (I_r - \hat{\beta}_1) \right|^{-(T-k+r)/2} &\leq |Q|^{T/2} |Q|^{-(T-k+r)/2} \\ &= |Q|^{(k-r)/2} \end{aligned} \quad (2.56)$$

So the integral of the product of these factors is bounded by  $\int |\alpha' D^{-1} \alpha|^{-(k-r)/2} d\alpha$ .

Hence, a sufficient condition for the existence of the posterior of  $\alpha$  is:

$$\int |\alpha' D^{-1} \alpha|^{-(k-r)/2} d\alpha < \infty. \quad (2.57)$$

The integrand has an asymptote at  $\alpha = 0_{(k \times r)}$ . We next analyze two shape features: an asymptote in the interior and tail behavior when  $\alpha$  tends to infinity. We shall show that the determinant in (2.57) is integrable around  $\alpha = 0$  despite the asymptote at  $\alpha = 0_{(k \times r)}$  and we derive conditions for the integrability of the tails.

## 2-dimensional vector case $r = 1, k = 2$

For simplicity, consider the integral on a ball  $A_k$  with radius  $R$  for the special case,  $k = 2$ ,  $r = 1$  where for ease of exposition we assume that the data matrices have been scaled and rotated such that  $Y_{-1}' Y_{-1} = I_k$ :

$$\int_{A_k} |\alpha' \alpha|^{-(k-r)/2} d\alpha = \iint_{\alpha_1^2 + \alpha_2^2 \leq R^2} (\alpha_1^2 + \alpha_2^2)^{-1/2} d\alpha_1 d\alpha_2. \quad (2.58)$$

We perform a polar coordinate transformation of  $\alpha_1, \alpha_2$  to show that the above integral is finite but depends on the value of  $R$ . Consider the change of variables:

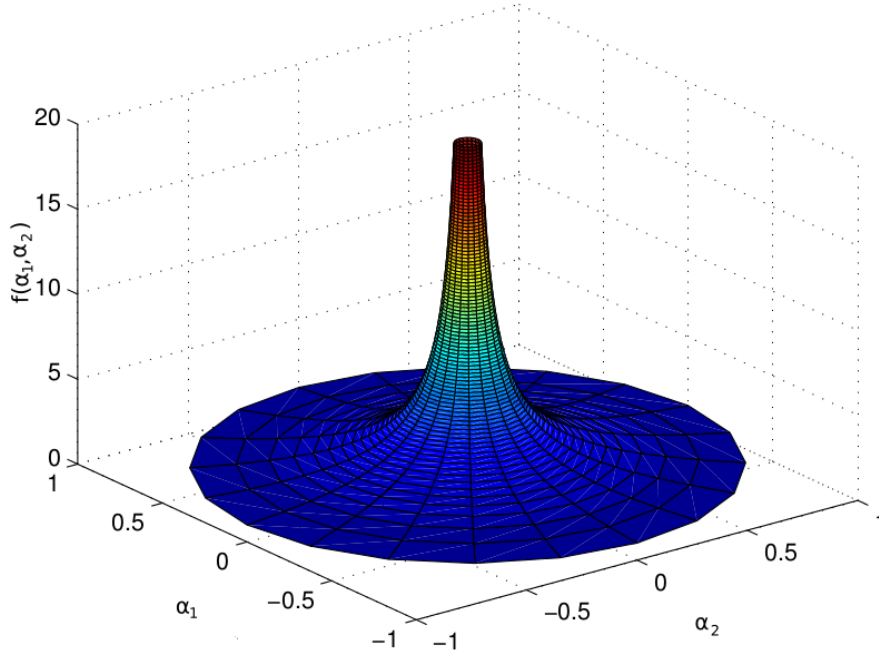
$$\begin{aligned} \alpha_1 &= \lambda \cos \theta, \quad \alpha_2 = \lambda \sin \theta \\ \lambda^2 &= \alpha_1^2 + \alpha_2^2, \quad \theta = \tan^{-1}(\alpha_2/\alpha_1), \end{aligned}$$

where  $\theta \in (0, 2\pi]$ ,  $\lambda > 0$  and the determinant of the Jacobian for this change of variables is

$$|J| = \begin{vmatrix} \cos \theta & -\lambda \sin \theta \\ \sin \theta & \lambda \cos \theta \end{vmatrix} = \lambda(\cos^2 \theta + \sin^2 \theta) = \lambda. \quad (2.59)$$

With the change of variables, the integral in (2.58) becomes:

$$\int_{\theta=0}^{2\pi} \int_{\lambda=0}^R (\lambda^2)^{-1/2} \lambda d\lambda d\theta = \int_{\theta=0}^{2\pi} \int_{\lambda=0}^R 1 d\lambda d\theta = 2\pi R, \quad (2.60)$$

Figure 2.1:  $f(\alpha) = (\alpha'\alpha)^{-1/2}$  for  $\alpha'\alpha \leq 1$ .

The integral corresponds to the volume under the graph of  $f(\alpha) = (\alpha'\alpha)^{-1/2}$ . The volume over the region  $\{\alpha | \alpha'\alpha \leq 1\}$  can be computed by integrating the surfaces of circles with radius  $f(\alpha)$  for  $1 \leq f(\alpha) < \infty$  and the surfaces  $\alpha$  of circles with radius 1 for  $0 \leq f(\alpha) < 1$ . Figure 2.1 illustrates this: for each function value  $f(\alpha) = (\alpha'\alpha)^{-1/2}$  with  $f(\alpha)$  as the horizontal ‘slice’ through the graph is a circle with radius  $1/f(\alpha)$ .

For any finite  $R$  the integral is bounded from which we conclude that the asymptote poses no problems. If however  $R$  tends to  $\infty$  the integral in (2.60) also goes to  $\infty$  at a rate  $R$ .

### General vector case $r = 1$

Next, we consider the case of  $r = 1$  but we relax the restriction  $k = 2$ . First we focus on the parameter space around the origin (where the asymptote is located). Regard  $h(\alpha'\alpha) = |\alpha'\alpha|^{-(k-r)/2} = (\alpha'\alpha)^{-(k-1)/2}$  as the kernel of a spherical density. Following e.g. theorems 1.5.5 and 2.1.3 from Muirhead (1982) this can be transformed to polar coordinates as

$$\begin{aligned}
 \alpha_1 &= \lambda \cos \theta_1 \\
 \alpha_2 &= \lambda \sin \theta_1 \cos \theta_2 \\
 \alpha_3 &= \lambda \sin \theta_1 \sin \theta_2 \cos \theta_3 \\
 &\vdots \\
 \alpha_{k-1} &= \lambda \sin \theta_1 \sin \theta_2 \dots \sin \theta_{k-2} \cos \theta_{k-1} \\
 \alpha_k &= \lambda \sin \theta_1 \sin \theta_2 \dots \sin \theta_{k-2} \sin \theta_{k-1}
 \end{aligned} \tag{2.61}$$

with  $\theta \in \Theta \subset \mathbb{R}^{k-1}$  with  $\Theta = \{\theta : \theta_{k-1} \in (0, 2\pi], \theta_i \in (0, \pi] \text{ for } i \neq k-1\}$  and  $\lambda > 0$ , such that  $\lambda^2 = \alpha' \alpha$  and the Jacobian is given by  $|J| = \lambda^{k-1} \prod_{i=1}^{k-2} \sin^{k-1-i} \theta_i$ . The  $\lambda, \theta_1, \dots, \theta_{k-1}$  are independent. All  $\theta_i$  have bounded density functions on a bounded support and can therefore be integrated out of the joint density resulting in a factor  $2\pi^{k/2}/\Gamma(k/2)$ . Hence the integral  $h(\alpha' \alpha)$  over a ball with radius  $R$  around the origin (where the asymptote is located) can be expressed as

$$\begin{aligned} \int_{\alpha' \alpha \leq R^2} h(\alpha' \alpha) d\alpha &= \int_{\lambda=0}^R \int_{\theta \in \Theta} h(\lambda^2) |J| d\theta d\lambda \\ &= \int_{\lambda=0}^R (\lambda^2)^{-(k-1)/2} \lambda^{k-1} \int_{\theta \in \Theta} \prod_{i=1}^{k-2} \sin^{k-1-i} \theta_i d\theta d\lambda \\ &= \frac{2\pi^{k/2}}{\Gamma(k/2)} \int_0^R 1 d\lambda = \frac{2\pi^{k/2} R}{\Gamma(k/2)}. \end{aligned} \quad (2.62)$$

Note that the existence of this expression does not depend on  $k$  and that it is equal to  $R$  times the surface area of a unit sphere in  $\mathbb{R}^k$ .

So also in the general vector case the asymptote poses no problems and for any finite  $R$  the integral is bounded. If however  $R$  tends to  $\infty$  the integral in (2.62) again goes to  $\infty$  at a rate  $R$ . So under diffuse priors in the linear normalization  $\beta_1 = I_r$  the tails of the marginal posterior are not integrable in case  $r = 1$  where  $\alpha$  reduces to a vector. We conjecture that since the integrand in terms of polar coordinates is constant, that any proper prior will suffice to 'repair' the integrability of the tails, for instance  $\alpha \sim N(0, cI)$  for  $c$  a large scalar. In combination with the integrability of the asymptote that we established this will result in a proper posterior.

### Matrix case

For the analysis of the asymptote in the matrix case we can use the transformation between  $\alpha$  and its singular value decomposition  $\alpha = USV'$  where  $U$  is a  $k \times r$  semi-orthogonal matrix with  $U'U = I_r$ ,  $V$  is a orthogonal  $r \times r$  matrix with  $V'V = I_r$  and  $S$  is a  $r \times r$  diagonal matrix with  $\lambda_i$ ,  $i = 1 \dots r$ , as diagonal elements. The  $\lambda_i$ 's denote the singular values in descending order, that is  $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_r \geq 0$ . Rennie (2006) shows using wedge product notation that the Jacobian of the transformation is proportional to

$$|J| \propto \prod_{i < j \leq r} (\lambda_i^2 - \lambda_j^2) \prod_{i \leq r} \lambda_i^{k-r}. \quad (2.63)$$

up to volume elements<sup>2</sup> of both the Stiefel manifold  $\mathbb{V}_{k,r} = \{U \in \mathbb{R}^{k \times r} : U'U = I_r\}$  related to  $U$  and the orthogonal group  $\mathbb{O}_r = \mathbb{V}_{r,r} = \{V \in \mathbb{R}^{r \times r} : V'V = I_r\}$  related to  $V$ . Note that the singular values of  $\alpha$  by definition are equal to the square roots of the eigenvalues of  $\alpha' \alpha$  and hence its determinant occurring in the integrand is equal to the product of the squared singular values  $\lambda_i$ , that is  $|\alpha' \alpha| = \prod_{i=1}^r \lambda_i^2$ .

<sup>2</sup>For ease of exposition we use this slightly less formal notation omitting wedge products in the intermediate step at the right hand side at the top of (2.64). See e.g. Muirhead (1982) or Rennie (2006) for formal wedge product notation.

Finally, we consider the area of integration  $\|\alpha'\alpha\|_2 \leq R$  around the asymptote where  $\|\cdot\|_2$  denotes the spectral norm which by definition equals the largest singular value (which is  $\lambda_1$  in our case), that is  $\lambda \in \Lambda_R = \{\lambda \in \mathbb{R}^r : \lambda : 0 \leq \lambda_r \leq \lambda_{r-1} \leq \lambda_1 \leq R\}$ . Note that this is a consistent generalization of the restriction  $\alpha'\alpha \leq R^2$  in terms of the Euclidean dot product in the vector case. Note also that the Frobenius norm (square root of the sum of the squared elements which equals the sum of the singular values) would be an equally valid generalization. The integral can thus be expressed as

$$\begin{aligned}
\int_{\|\alpha\|_2 \leq R} |\alpha'\alpha|^{-(k-r)/2} d\alpha &= 2^{-r} \int_{U'U=I} \int_{V'V=I} \int_{\Lambda_R} \prod_{i=1}^r (\lambda_i^2)^{-(k-r)/2} |J| d\lambda dV dU \\
&= 2^{-r} \text{Vol}(\mathbb{O}_r) \text{Vol}(\mathbb{V}_{k,r}) \int_{\Lambda_R} \prod_{i < j \leq r} (\lambda_i^2 - \lambda_j^2) d\lambda \\
&= \frac{2^r \pi^{r(k+r)/2}}{\Gamma_r(k/2) \Gamma_r(r/2)} \int_{\Lambda_R} \prod_{i < j \leq r} (\lambda_i^2 - \lambda_j^2) d\lambda \tag{2.64}
\end{aligned}$$

using the fact that the volume of the Stiefel manifolds is given by  $\text{Vol}(\mathbb{V}_{k,r}) = \frac{2^r \pi^{kr/2}}{\Gamma_r(k/2)}$  and  $\text{Vol}(\mathbb{O}_r) = \text{Vol}(\mathbb{V}_{r,r}) = \frac{2^r \pi^{r^2/2}}{\Gamma_r(r/2)}$  and the factor  $2^{-r}$  arises because of the uniqueness of the singular value decomposition up to simultaneous sign changes of corresponding columns of  $U$  and  $V$  which could be enforced for instance by imposing a positive sign for the first nonzero entry in each column of  $U$ . In the special case  $r = 1$  the product in the integrand is empty and the integrand becomes equal to 1, and therefore (2.64) reduces to (2.62).

The integrand is a polynomial in the  $\lambda_i$ 's and the area of integration is bounded. Hence, we conclude the integral over this bounded region is finite despite the fact that it contains an asymptote at  $|\alpha'\alpha| = 0$ .

For the analysis of the tail behaviour of the integral we use a slightly different approach than in the vector case. For any  $\varepsilon > 0$  consider the area  $\alpha'\alpha \succ \varepsilon I_r$ , that is  $\{\alpha : \alpha'\alpha - \varepsilon I_r \text{ is positive definite}\}$ . Since  $\alpha'\alpha - \varepsilon I_r \succ 0$  we have by adding and subtracting  $\alpha'\alpha$  that also  $(\alpha'\alpha + \alpha'\alpha) - (\varepsilon I_r + \alpha'\alpha) \succ 0$ . This implies that  $|\alpha'\alpha + \alpha'\alpha| > |\varepsilon I_r + \alpha'\alpha|$ . We use this result to show the following inequality:

$$|\alpha'\alpha| = 2^{-r} |\alpha'\alpha + \alpha'\alpha| > 2^{-r} |\varepsilon I_r + \alpha'\alpha| \tag{2.65}$$

On the other hand we have that  $|\alpha'\alpha| < |\varepsilon I_r + \alpha'\alpha|$  so

$$2^{-r} |\varepsilon I_r + \alpha'\alpha| < |\alpha'\alpha| < |\varepsilon I_r + \alpha'\alpha| \tag{2.66}$$

Our function is 'sandwiched' between two expressions which only differ by a factor  $2^{-r}$  which does not depend on  $\alpha$ . We note that both these expressions are proportional to  $p_{Mt}(\alpha|I_k, \varepsilon I_r, k-r)$ , so the integrability of  $|\alpha'\alpha|^{-(k-r)/2}$  over the region  $\alpha'\alpha \succ \varepsilon I_r$  is determined by the integrability of  $p_{Mt}(\alpha|I_k, \varepsilon I_r, k-r)$ . So

$$\begin{aligned}
\int_{\alpha'\alpha \succ \varepsilon I_r} |\alpha'\alpha|^{-(k-r)/2} d\alpha &< 2^{-r} \int_{\alpha'\alpha \succ \varepsilon I_r} |\varepsilon I_r + \alpha'\alpha|^{-(k-r)/2} d\alpha \\
&\propto 2^{-r} \int_{\alpha'\alpha \succ \varepsilon I_r} p_{Mt}(\alpha|I_k, \varepsilon I_r, k-r) d\alpha. \tag{2.67}
\end{aligned}$$

By using properties of the matrix-variate  $t$  distribution the latter is integrable only if  $r = 0$  because the condition mentioned below (2.109) requires that  $k - r > k + r - 1$  or  $r < 1/2$ .

**Proposition IV** *Given the standard form of a cointegration model under linear normalization and using a diffuse class of priors, the marginal posterior distribution of  $\alpha$ , with density (2.44) is integrable on any finite area even despite the fact that it has an asymptote at  $|\alpha'\alpha| = 0$ . The tails however are too heavy to yield a proper posterior when  $r > 0$ .*

It is thus necessary to either use at least a weakly informative prior such as a normal prior  $N(0, cI)$  with  $c$  a large constant which will be sufficient to overcome the nonintegrability of the tails or alternatively one could bound the support of the parameters to a finite region.

## 2.5 Regularization through area restrictions: prior under orthogonal normalization

Given a diffuse prior and under linear normalization we have shown that the marginal posteriors of the parameters of interest of a standard cointegration model are not regular in the sense that they do not belong to a known class of densities like the matrix-variate  $t$  densities. This latter property holds only in the extreme case when all data are stationary. Also in the case when all series are random walks one encounters regular posteriors. For cases where linear combinations of time series are stationary, we have shown that the irregularity of the posteriors does not occur in the interior of the parameter region. There is, however, an existence problem of the posteriors with zero-th or higher order moments when the cointegration and, in particular, the adjustment parameters tend to infinity. It is therefore natural to explore an approach where weak regularizing prior information is introduced that makes use of restrictions, in particular, plausible restrictions on the range of the parameters. That is a line of research that is pursued in the present section.

### 2.5.1 Alternative specification and identification

In general an  $n \times k$  matrix of rank  $r$  has  $(n + k)r - r^2$  free elements, that is  $(n - r)(k - r)$  restrictions. In our case, the  $k \times k$  matrix  $\Pi$  has rank  $r$  and therefore it has  $2kr - r^2$  independent free elements and  $(k - r)^2$  restrictions. The matrices  $\alpha$  and  $\beta$  in the parametrization  $\Pi = \beta\alpha'$  with  $\text{rank}(\Pi) = r$  together have  $2kr$  elements, which are  $r^2$  too many to identify  $\alpha$  and  $\beta$ . The normalization  $\beta_1 = I_r$  that we used in the previous sections exactly accounts for the additional  $r^2$  required restrictions. The parametrization  $\Pi = \beta\alpha'$  can be linked to the singular value decomposition  $\Pi = USV'$ , where the rectangular  $k \times r$  matrix  $U$  is an element of the Stiefel manifold  $U'U = I_r$  and the square  $r \times r$  matrix  $V$  is an element of the manifold of orthogonal matrices  $V'V = I_r$ .  $S$  is a diagonal  $r \times r$  matrix with positive diagonal entries equal to the singular values of  $\Pi$ . We denote the vector of these diagonal elements as  $\lambda = (\lambda_1, \dots, \lambda_r)'$ . Note that the manifolds on which  $U$  and  $V$

are defined have finite volume. The manifold on which  $\lambda$  is defined is not bounded and we shall come back to that later.

E.g. Kleibergen and Van Dijk (1998) and Kleibergen and Paap (2002) explicitly link their parametrization to the singular value decomposition and they combine it with the linear restriction  $\beta_1 = I_r$ . This linear normalization subsequently implies a mapping from these manifolds to cartesian coordinates in Euclidean space, that is  $\alpha \in \mathbb{R}^{k \times r}$  and  $\beta_2 \in \mathbb{R}^{(k-r) \times r}$ . This mapping thus transforms from manifolds with finite volume (except  $\lambda$ ) to unbounded spaces.

Another common normalization of  $\beta$  used in the literature is  $\beta'\beta = I_r$ . A major motivation for the choice of this orthogonal normalization of the matrix  $\beta$  is that in this case no preferred ordering of the variables is imposed and the region of integration for  $\beta$  is bounded. In the case of a VAR these may be reasonable assumptions in several situations, in particular, when one considers a set of similar price indices or quantity series.

We emphasize that this normalization alone is not sufficient to identify both  $\alpha$  and  $\beta$ . This normalization imposes only  $r(r+1)/2$  unique restrictions, because of the symmetry of  $\beta'\beta$ , so an additional  $r(r-1)/2$  restrictions are required. One could impose these on  $\beta$  but this should be done with caution in order to avoid the issue of imposing too much structure through the combination of ordering, restricting and assigning a flat prior.

In the remainder of this section we propose an approach that more directly uses the structure of the singular value decomposition.

As specified above, the singular value decomposition is not uniquely defined. Any simultaneous permutation of the columns of  $U$ ,  $S$  and  $V$  also constitutes a singular value decomposition. A common way to avoid this ambiguity is by ordering the singular values that occur on the diagonal of  $S$  as  $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_r \geq 0$ . We shall use this ordering. Ordering the singular values is also more straightforward than devising an ordering of the columns of  $U$  and  $V$  directly (or the columns of  $\alpha$  and  $\beta$  for that matter).

Because of this ordering each element  $\lambda_{i+1}$  for  $i = 1, \dots, r-1$  is bounded by  $\lambda_i$ . Only  $\lambda_1$  remains unbounded towards  $+\infty$ . Integrability is thus determined by the behaviour of  $\lambda_1$ .

Having fixed the ordering of the singular values the uniqueness of the singular value decomposition when all  $\lambda_i$ 's are different is up to simultaneous sign changes of corresponding columns of  $U$  and  $V$  which could be mitigated for instance by imposing a positive sign for the first nonzero entry in each column of  $U$ . Finally, if a singular value occurs more than once, then the columns of  $U$  and  $V$  corresponding to these singular values are not uniquely defined. Any other orthonormal basis that spans the same space will also do. Although in this particular case the transformation between the matrix  $\Pi$  and its singular value decomposition  $(U, S, V)$  is still not invertible everywhere, this is however an event with zero measure and we observe that the Jacobian in (2.63) of this transformation equals 0 whenever a repeated singular value occurs because then the factor  $\lambda_i^2 - \lambda_j^2$  will be 0 for some  $i < j$ .

We analyse the specification in which we combine  $\beta'\beta = I_r$  with  $\alpha'\alpha = I_r$  in the parametrization  $\Pi = \beta\Lambda\alpha'$  with  $\Lambda$  diagonal. This corresponds directly to the singular value decomposition  $\Pi = USV'$  with  $\beta = U$ ,  $\alpha = V$  and  $\Lambda = S = \text{diag}(\lambda)$ . The restriction  $\alpha'\alpha = I_r$  imposes  $r(r+1)/2$  restrictions which amounts to  $r$  restrictions more than required, but  $\lambda$  subsequently provides these extra  $r$  degrees of freedom.



$\Lambda$  and  $\alpha$  in this parametrization combine into  $\alpha$  in the usual parametrization  $\Pi = \beta\alpha'$  as in the previous bullet.

The advantage of this specification is that now both  $\alpha$  and  $\beta$  have finite support, which helps to avoid the issues with integrability of the tails as encountered in the previous section. If difficulties arise it will be in the parameter  $\lambda$ , and if so it is also clear they will also have to be repaired in  $\lambda$ .

Regarding the econometric interpretation of the parametrization  $\Pi = \beta\Lambda\alpha'$  we may think of  $\beta'y_t$  as the deviation from the  $r$  cointegrating relations  $\beta'y_t = 0$  between the  $k$  variables  $y_t$ , which is similar to the role of  $\beta$  in the more usual parametrization  $\Pi = \beta\alpha'$ . The interpretation of  $\lambda$  is that of the rate of adjustment of the system towards each of the  $r$  cointegrating relations.  $\alpha$  in the parametrization  $\Pi = \beta\Lambda\alpha'$  describes the contribution of each of the  $k$  variables  $y_t$  to the adjustment towards each of these  $r$  cointegrating relations. This has advantages over the more usual parametrization  $\Pi = \beta\alpha'$  in which the speed of adjustment towards the cointegrating relations is amalgamated with the distribution of these adjustments over the variables into one single parameter matrix (also denoted  $\alpha$ ).

Each data vector  $y_t$  defines a vector in  $k$ -dimensional space. The geometric interpretation is that  $\beta$  defines  $r$  directions in the space of the data.  $\Lambda$  scales in these directions and  $\alpha$  rotates the result to a  $r$  dimensional subspace of the data.

To distinguish the parameter matrix  $\alpha$  in  $\Pi = \beta\Lambda\alpha'$  from the parameter matrix  $\alpha$  in the usual parametrization we shall denote the latter by  $\alpha^*$  such that  $\Pi = \beta\alpha^{*'} in the remainder of this section. In order to translate results on  $\alpha$  and  $\Pi = \beta\Lambda\alpha'$  back and forth to  $\alpha^*$  and  $\Pi = \beta\alpha^{*'} we now briefly describe how they are related. Both parameterizations are linked by the relation  $\alpha^* = \alpha\Lambda$ . This can be seen when we combine  $\beta'\beta = I_r$  with  $\alpha^{*'}\alpha^* = S$  in the parametrization  $\Pi = \beta\alpha^{*}'$  where  $S$  is a  $r \times r$  diagonal matrix with  $\lambda_i$ ,  $i = 1, \dots, r$ , as diagonal elements. The relation with the singular value decomposition  $\Pi = USV'$  is  $\beta = U$ ,  $\alpha^* = VS = \alpha\Lambda$ . This also gives exactly the number of required restrictions: all off-diagonal elements of  $\alpha^{*'}\alpha^*$  are constrained to 0 and because of the symmetry of  $\alpha^{*'}\alpha^*$  each off-diagonal element occurs twice which results in  $r(r-1)/2$  unique restrictions. In terms of the columns  $\alpha_i^*$  of  $\alpha^*$ :  $\alpha_i^{*'}\alpha_i^* = \lambda_i^2$  for  $i = 1, \dots, r$  and  $\alpha_i^{*'}\alpha_j^* = 0$  for  $i \neq j$ .$$

## 2.5.2 Prior choice and posterior moments

In the specification  $\Pi = \beta\Lambda\alpha'$  uniform priors can be specified for  $\alpha$  en  $\beta$  on their respective Stiefel manifolds. For  $\Sigma$  we again specify a diffuse prior and we assume all marginal prior to be independent, that is

$$p(\beta, \alpha, \lambda, \Sigma) = p(\beta)p(\alpha)p(\lambda)p(\Sigma) \quad (2.68)$$

with

$$p(\alpha) \propto \begin{cases} 1 & \text{if } \alpha'\alpha = I_r, \\ 0 & \text{otherwise} \end{cases} \quad (2.69)$$

$$p(\beta) \propto \begin{cases} 1 & \text{if } \beta'\beta = I_r, \\ 0 & \text{otherwise} \end{cases} \quad (2.70)$$

$$p(\Sigma) = |\Sigma|^{-h/2} \quad \text{if } \Sigma \text{ is symmetric and positive definite} \quad (2.71)$$

and we use again the specific case of  $h = k + 1$ . We now discuss the choice of the prior on the singular values  $\lambda$  in more detail when we explore the integrability of the posterior in relation to this prior  $p(\lambda)$ .

Due to the similarity in the prior and the likelihood, (2.13)–(2.14), with the distinction (i) that all elements of  $\beta$ , not only elements of  $\beta_2$ , are now random variables, (ii) that we now have  $\alpha^* = \alpha\Lambda$  instead of  $\alpha$  and (iii) that we now include the prior  $p(\lambda)$  (which is independent of the priors on the other parameters), we can write

$$\begin{aligned} p(\alpha, \beta, \lambda) &\propto |(\Delta Y - Y_{-1}\beta\Lambda\alpha')'(\Delta Y - Y_{-1}\beta\Lambda\alpha')|^{-T/2} p(\lambda) \\ &= |\Delta Y' M_{Y_{-1}\beta\Lambda} \Delta Y + (\alpha - \hat{\alpha}) (\Lambda\beta' Y_{-1}' Y_{-1} \beta \Lambda) (\alpha - \hat{\alpha})'|^{-T/2} p(\lambda) \end{aligned} \quad (2.72)$$

When we integrate this posterior with respect to  $\alpha$  over the manifold  $\alpha'\alpha = I_r$  we can derive the following bound:

$$\begin{aligned} &\int_{\alpha'\alpha=I} p(\alpha, \beta, \lambda|Y) d\alpha \propto \\ &= \int_{\alpha'\alpha=I} |\Delta Y' M_{Y_{-1}\beta\Lambda} \Delta Y + (\alpha - \hat{\alpha}) (\Lambda\beta' Y_{-1}' Y_{-1} \beta \Lambda) (\alpha - \hat{\alpha})'|^{-T/2} p(\lambda) d\alpha \\ &\leq \int_{\alpha'\alpha=I} |\Delta Y' M_{Y_{-1}\beta\Lambda} \Delta Y|^{-T/2} p(\lambda) d\alpha \\ &= |\Delta Y' M_{Y_{-1}\beta\Lambda} \Delta Y|^{-T/2} p(\lambda) \int_{\alpha'\alpha=I} d\alpha \\ &\propto \left( \frac{|\Lambda\beta' Y_{-1}' M_{\Delta Y} Y_{-1} \beta \Lambda|}{|\Lambda\beta' Y_{-1}' Y_{-1} \beta \Lambda|} \right)^{-T/2} p(\lambda) \text{Vol}(\mathbb{V}_{k,r}) \\ &= \left( \frac{|\beta' Y_{-1}' M_{\Delta Y} Y_{-1} \beta|}{|\beta' Y_{-1}' Y_{-1} \beta|} \right)^{-T/2} p(\lambda) \text{Vol}(\mathbb{V}_{k,r}), \end{aligned} \quad (2.73)$$

where the second to last step follows from (2.101) and the last step follows since both numerator and denominator are of the form  $|\Lambda X \Lambda|$  for some matrix  $X$  and can be written as  $|\Lambda||X||\Lambda|$  such that the factors  $|\Lambda|$  in numerator and denominator cancel against each other.

The first factor in (2.73) is bounded by products of eigenvalues similar to the previous section. The volume  $\text{Vol}(\mathbb{V}_{k,r})$  of the Stiefel manifold of  $k \times r$  orthogonal matrices  $\alpha$  is a finite constant. So integrability of (2.72) depends on  $p(\lambda)$ .

The alternative route via integrating over  $\beta$  proceeds as follows:

$$\begin{aligned}
& \int_{\beta' \beta = I} p(\alpha, \beta, \lambda | Y) d\beta \propto \\
& = \int_{\beta' \beta = I} \left| (Y'_{-1} M_{\Delta Y \alpha_{\perp}} Y_{-1})^{-1} + (\beta - \hat{\beta} \Lambda^{-1}) \Lambda \alpha' D^{-1} \alpha \Lambda (\beta - \hat{\beta} \Lambda^{-1})' \right|^{-T/2} p(\lambda) d\beta \\
& \leq \int_{\beta' \beta = I} \left| (Y'_{-1} M_{\Delta Y \alpha_{\perp}} Y_{-1})^{-1} \right|^{-T/2} p(\lambda) d\beta \\
& = \left| (Y'_{-1} M_{\Delta Y \alpha_{\perp}} Y_{-1})^{-1} \right|^{-T/2} p(\lambda) \int_{\beta' \beta = I} d\beta \\
& \propto \left( \frac{|\alpha'_{\perp} \Delta Y' M_{Y_{-1}} \Delta Y \alpha_{\perp}|}{|\alpha'_{\perp} \Delta Y' \Delta Y \alpha_{\perp}|} \right)^{-T/2} p(\lambda) \text{Vol}(\mathbb{V}_{k,r}), \tag{2.74}
\end{aligned}$$

where the last step follows again from (2.101). The first factor, which is a function of  $\alpha$ , is bounded by products of eigenvalues, and again integrability of (2.74) depends on  $p(\lambda)$ .

As a starting point for the specification of an uninformative prior, suppose that we specify a diffuse prior on  $\Pi$  on the manifold of  $k \times k$  matrices with rank  $r$ , that is

$$p(\Pi) \propto 1 \text{ if } \text{rank}(\Pi) = r, 0 \text{ otherwise} \tag{2.75}$$

then using the Jacobian of the transformation  $\Pi = \beta \Lambda(\lambda) \alpha'$  which is similar to (2.63) we obtain that the implied prior for  $(\alpha, \beta, \lambda)$  equals

$$p(\alpha, \beta, \lambda) \propto p(\Pi(\alpha, \beta, \lambda)) \left| \frac{\partial \text{vec}(\Pi(\alpha, \beta, \lambda))}{\partial \text{vec}(\alpha, \beta, \lambda)'} \right| \tag{2.76}$$

$$\propto \prod_{i < j \leq r} (\lambda_i^2 - \lambda_j^2) \prod_{i \leq r} \lambda_i^{k-r}. \tag{2.77}$$

This implies independent priors on  $\alpha$ ,  $\beta$  and  $\lambda$  with  $\alpha$  and  $\beta$  uniform similar to (2.69) and (2.70). However, the implied prior on the singular values

$$p(\lambda) = \prod_{i < j \leq r} (\lambda_i^2 - \lambda_j^2) \prod_{i \leq r} \lambda_i^{k-r} \tag{2.78}$$

is not integrable as  $\lambda_i \rightarrow \infty$ . The factor  $\prod_{i < j \leq r} (\lambda_i^2 - \lambda_j^2)$  in (2.78) results from the ordering of the singular values that we assume in the singular value decomposition and regularizes the posterior by letting the prior go to 0 whenever two (or more) singular values  $\lambda_i$  and  $\lambda_j$  for  $i \neq j$  are equal, because in that case the factor  $\lambda_i^2 - \lambda_j^2$  will equal 0.

We note that the other factor  $\prod_{i \leq r} \lambda_i^{k-r}$  in (2.78) can be shown to correspond to the embedding prior of Kleibergen and Paap (2002) on  $(\alpha^*, \beta)$  conditional on the rank reduction under a flat prior on  $\Pi$ . Their prior is in that case given by

$$p(\beta, \alpha^*) \propto |\beta' \beta|^{(k-r)/2} |\alpha^* \alpha^*|^{(k-r)/2}. \tag{2.79}$$

We can rewrite their prior adapted for our normalization using  $\alpha^* = \alpha\Lambda$ ,  $\alpha'\alpha = I_r$  and  $\beta'\beta = I_r$  as

$$|\beta'\beta|^{(k-r)/2} |\alpha^*\alpha'|^{(k-r)/2} = |\beta'\beta|^{(k-r)/2} |\Lambda\alpha'\alpha\Lambda|^{(k-r)/2} \quad (2.80)$$

$$\begin{aligned} &= |I_r|^{(k-r)/2} |\Lambda I_r \Lambda|^{(k-r)/2} = |\Lambda\Lambda|^{(k-r)/2} \\ &= |\Lambda|^{k-r} = \prod_{i=1}^r \lambda_i^{k-r}. \end{aligned} \quad (2.81)$$

The connection with the previous section is that this prior regularizes the vertical asymptote at  $|\alpha^*\alpha'| = 0$  with  $\alpha^* = \alpha \text{diag}(\lambda)$ , but we found there that it is not the asymptote that leads to integrability problems but rather the tails. This is also the case here for  $\lambda_1 \rightarrow \infty$  whenever the product of the remaining singular values, that is  $\prod_{i=2}^r \lambda_i$ , is non-zero.

We now try to specify an uninformative or weakly informative prior on  $\lambda$  using a more direct approach. Initially we disregard the ordering of singular values and we could then use the following approach. Since  $\lambda_i > 0$  specifying a diffuse prior on  $\log \lambda_i$  would correspond to  $p(\lambda_i) \propto \lambda_i^{-1}$  which is analogous to a diffuse prior  $p(\sigma^2) \propto \sigma^{-2}$  for a variance parameter  $\sigma^2$ . In this case the prior for the vector  $\lambda$  equals  $p(\lambda) \propto \prod_{i=1}^r \lambda_i^{-1}$ . Note that  $\lambda_i$  equals  $(\alpha_i^*\alpha_i^*)^{1/2}$  such that both correspond to the singular values of  $\Pi$ . The implied prior in the specification  $\Pi = \beta\alpha^*$  is thus given by  $p(\alpha^*) = \prod_{i=1}^r (\alpha_i^*\alpha_i^*)^{-1/2} = |\alpha^*\alpha'|^{-1/2}$ .

If we also include the ordering of the singular values  $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_r \geq 0$  in the prior specification we first note that the singular values  $\lambda_i$  for all  $i > 1$  are bounded by  $\lambda_{i-1}$ . So given  $\lambda_1$  the other  $\lambda_i$  are jointly restricted to a bounded (hyper-)triangular region  $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_r \geq 0$ . Conditional on  $\lambda_1$  we specify a joint uniform prior on  $(\lambda_2, \dots, \lambda_r)$  on this support. Only the largest singular value  $\lambda_1$  has infinite support and requires a prior of which the tails go to zero fast enough.

We now use the connection to a Dirichlet distribution to find a prior for  $\lambda_1$  that is consistent with the other  $\lambda_i$ . In order to do this we transform the  $\lambda_i$  into the increments  $\delta_i$  as follows:

$$\delta_i = \lambda_i - \lambda_{i+1}, \text{ for } i = 1, \dots, r \quad (2.82)$$

with inverse transformation  $\lambda_i = \sum_{j=i}^r \delta_j$ . Its Jacobian is given by

$$\left| \frac{\partial \lambda}{\partial \delta'} \right| = \begin{vmatrix} 1 & 1 & \dots & 1 \\ 0 & 1 & \dots & 1 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \end{vmatrix} = 1, \quad (2.83)$$

which means that we can easily transform from  $\lambda$  to  $\delta$  and vice versa. In particular a joint uniform distribution on  $\lambda_i$  also implies a joint uniform distribution on  $\delta_i$  over the simplex on which it is defined, that is  $\delta_i > 0$  and  $\sum_{i=1}^r \delta_i = \lambda_1$ .

A Dirichlet distribution  $D(1, \dots, 1)$  also corresponds to a uniform distribution on a simplex. This Dirichlet distribution can be constructed from  $r$  i.i.d. random variables from an exponential distribution with any rate  $\theta > 0$ . Let  $p(\delta_i) \sim \exp(-\delta_i\theta)$ . Then

$$\frac{\delta_i}{\sum_{i=1}^r \delta_i} = \frac{\delta_i}{\lambda_1} \sim D(1, \dots, 1), \quad (2.84)$$

as required. We now also can find a prior for  $\lambda_1$  that is fully consistent with the joint uniform prior on the  $\lambda_i$  for  $i > 1$  on its support since all  $\lambda_i$  are derived from the same i.i.d. joint distribution of the increments  $\delta_i$ :

$$\lambda_1 = \sum_{j=1}^r \delta_j \sim \text{Gamma}(r, \theta) \quad (2.85)$$

and its density is thus given by

$$p(\lambda_1) \propto \lambda_1^{r-1} \exp(-\lambda_1 \theta). \quad (2.86)$$

The density of the uniform prior  $p(\lambda_2, \dots, \lambda_r | \lambda_1)$  equals the inverse of the volume of the simplex on which it is defined. The volume equals  $\lambda_1^{r-1} / (r-1)!$  where the factor  $1/(r-1)!$  results from the Jacobian of the transformation from (unit) box to (unit) simplex.

So the joint prior of  $\lambda$  on its support  $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_r \geq 0$  in this approach is thus given by

$$p(\lambda) = p(\lambda_2, \dots, \lambda_r | \lambda_1) p(\lambda_1) \propto \exp(-\lambda_1 \theta). \quad (2.87)$$

We can summarize the results from this section as follows. Using the parametrization  $\Pi = \beta \Lambda \alpha'$  and the normalizing restrictions  $\alpha' \alpha = I_r$ ,  $\beta' \beta = I_r$  and  $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_r \geq 0$  all parameters except  $\lambda_1$  are defined on bounded sets. An natural choice for an uninformative prior is the uniform prior over these sets. Only  $\lambda_1$  is defined on an infinite interval. A natural choice for  $\lambda_1$  that is consistent with the uniform prior on the simplex for  $\lambda_2, \dots, \lambda_r | \lambda_1$  is the exponential distribution. Another way to look at this, is that although  $\lambda \in [0, \infty)$  has infinite support, it can also be transformed to the unit interval on which a uniform prior can be specified. By doing so, all model parameters (except the covariance matrix  $\Sigma$ ) are bounded to finite areas. Specifically, when either the transformation  $\lambda^b = \exp(-\lambda) \in [0, 1)$  or  $\lambda^\# = 1 - \exp(-\lambda) \in [0, 1)$  is used and a standard uniform density is specified on  $\lambda^b$  or  $\lambda^\#$  then  $\lambda$  also has a standard exponential distribution. Using a similar argument the rate parameter  $\theta$  could be included by specifying a uniform prior on e.g.  $\exp(-\theta \lambda)$ . A final remark concerns the rate  $\theta$  of the exponential distribution. By choosing  $\theta$  to a value close to 0, the exponential distribution tends towards a flat distribution over the positive real numbers.

Note that also in Chapter 4 we shall also employ Dirichlet priors and uniform priors defined on a simplex when in order to deal with proportions (of total variance) and for partitioning a bounded interval (in that case of frequencies of cyclical components) into an ordered set of parameters.

## 2.6 Conclusions and remarks

We have provided an analysis of the shape and existence of the posterior probability distribution function of the parameters of interest in a cointegration model. These parameters are divided into two sets: cointegration parameters that refer to long run stable patterns in the series and adjustment parameters that refer to error correction mechanisms in the model. The nonstandard shape of the joint and marginal posteriors

refers to two issues: the joint posteriors may have a ridge in the surface, which implies that marginals become unbounded in the interior of the region of integration but we have shown that the asymptote is integrable. Secondly, tails of posteriors may be too heavy in the sense of not tending to zero sufficiently fast. We note that the conditionals belong to a regular class of  $t$  densities.

We present specific conclusions by distinguishing between three typical cases of the behavior of economic time series analyzed through a cointegration model with linear normalization:

*All time series are stationary:  $k = r$*

Since the matrix  $\beta$  is the identity matrix, ( $\beta_2$  is zero), the number of variables is equal to the number cointegrating relations, that is  $k = r$ . The matrix  $\alpha$  is equal to the matrix  $\Pi$  of unrestricted parameters and one is in the situation of the multivariate regression framework where the marginal posterior of  $\Pi$  is a matrixvariate  $t$  density with parameters that only depend on given data.

*All time series are random walks:  $r = 0$*

In this situation there is no error correction or adjustment process. Thus the matrix  $\alpha$  is zero and  $r = 0$ . In the basic model there are no uncertain coefficients in the equation but only a covariance matrix of the disturbances. One is again in a case that can be handled in a regular way.

*All time series are random walks and some combinations are stationary:  $0 < r < k$*

The marginal posterior of the cointegration parameters  $\beta_2$  exists but no first or higher order moments and this result is independent of the value of  $r$  or  $k$ . The marginal posterior of the adjustment parameters  $\alpha$  is integrable on any finite area including the area around an asymptote but not on the whole real space. Thus, for the well-known cointegration model of two time series with one cointegrating relation,  $k = 2$  and  $r = 1$ , it follows that the posterior of  $\alpha$  is proper on a finite region but not everywhere in real space.

A natural follow up of the line of research reported so far is to explore the use of so-called regularization priors which are weakly informative priors that modify the likelihood only slightly and make it more regular. Using plausible restrictions on the range of the parameters seems reasonable. The assumption that cointegration as well as adjustment parameters may take values in the whole real space is from an economic theory point of view not very plausible. In the present chapter we proposed alternative forms of normalization as restrictions which help to get more regular behavior. We emphasize that there exist several other related approaches in this context, see the survey by Koop et al. (2006). We refer to Strachan and Van Dijk (2003), Strachan and Van Dijk (2013) and Villani (2005) who investigate the case of a flat prior on the rank of the cointegration matrix. We refer also to Hamilton et al. (2007) for a survey and more background on the issue of normalization.

Another approach to regularization is the use of the class of Information Matrix priors. In situations where the likelihood tends to become flat the determinant of the Information Matrix tends to zero. Kleibergen and Van Dijk (1998) and Kleibergen and Paap (2002) introduce a Jeffreys' prior and the equivalent embedding prior in order to obtain a more regular shape of the posterior.

A third approach to make the tail behavior more regular is to make use of weakly normal priors, which is an exponential class of densities with thin tails. For the adjustment

parameters there exists the class of so-called Minnesota priors, see Doan et al. (1992) and for the cointegration relations there exists the class of normal priors on the long run which are intended to anchor the cointegration relations. A first paper was due to Schotman and Van Dijk (1991) and this is further extended by Villani (2009) and recently by Giannone et al. (2015).

It is of course also possible to make use of structural economic models like structural VAR models and DSGE models. Then one faces different problems like the credibility of structural restrictions, see Sims (1980). In this chapter the focus was on the information content in those economic time series that contain non-stationary as well as stationary components.

An other conclusion refers to the meaning of these theoretical results for the implementation of a basic method of Bayesian simulation: the Gibbs method. In case of linear normalization one encounters several model specification where the conditional posteriors exist but not the joint posterior. Application of the Gibbs sampler would give erroneous results in such situations. Further, the Gibbs sampler may be very inefficient when one is in a situation of near-unit roots. Restrictions on the range of the parameters like orthogonal normalizations allow for more regular posteriors but different Monte Carlo integration methods are then often more efficient than the Gibbs sample.

We end this chapter with two remarks.

First, we emphasize that there exists an interesting analogue between inference in a cointegration model and inference in an Instrumental Variable (IV) regression model (also called Incomplete Simultaneous Equations model, see Zellner, Bauwens, and Van Dijk (1988)). There exists a large literature on this topic. In the present chapter, we only make use of Zellner, Ando, Baştürk, Hoogerheide, and Van Dijk (2014) for this comparison.

In the cointegration model there is a *matrix* of cointegration parameters and one has a matricvariate  $t$  density while in the IV model one has a *vector* of endogenous parameters and a multivariate  $t$  density. In the cointegration model one needs to "correct" for many more degrees of freedom. Therefore, in the cointegration model, there is only existence of the zero-th moment of  $\beta_2$  and not a kind of analogue to the overidentification result in an IV model. This result extends the analysis of Kleibergen and Van Dijk (1994).

Next, one may compare the marginal posterior of the adjustment parameters, equation (2.51), with the marginal posterior of the instrumental variable parameters, equation (11) in Zellner, Ando, Baştürk, Hoogerheide, and Van Dijk (2014). In both cases there exists the result of the "integrable asymptote" in the interior but in Zellner, Ando, Baştürk, Hoogerheide, and Van Dijk (2014) one can "ignore" the tail behavior of the instrumental variable parameters since the matricvariate  $t$  density that is listed in the first factor of the product at the right hand side of equation (11) has such "good" tail behavior (tends to zero fast due to large degrees of freedom) that it "kills" the heavy tail of the third factor in the product. However, in the case of a cointegration model, there is no such penalty on the tails of the density of the adjustment parameters and we have a non-integrable posterior on the whole real space.

A second remark refers to our use of a flat prior which has been criticized for not being invariant to a parameter transformation. This is a correct argument and two well-known examples are that in a AR(2) model being uniform on the two lagged variables

does not imply that one is uniform on the eigenvalues. One maybe informative in an undesired way on the stability of the equation. Second, in a simple Keynesian income-consumption model with exogenous government expenditures a flat prior on the marginal propensity to consume parameter implies that one is very informative on the multiplier, again, in an undesired way. These examples indicate that a researcher has to be explicit on the parameters of interest and she may perform prior predictive analysis in order to investigate implications of flat priors of structural parameters for forecasting and policy scenarios. This is an interesting topic but outside the scope of the present chapter. For some examples, we refer to Van Dijk and Klok (1980) and Baştürk et al. (2014a).



## APPENDICES

### 2.A Some useful results on matrices and determinants

#### Linear regression

Linear regression:

$$(Y - X\beta)'(Y - X\beta) = Y'M_X Y + (\beta - \hat{\beta})'X'X(\beta - \hat{\beta}), \quad (2.88)$$

where  $\hat{\beta} = (X'X)^{-1}X'Y$  and  $M_X = I - X(X'X)^{-1}X$ .

Frisch-Waugh:

$$\begin{aligned} & (Y - X_1\beta_1 - X_2\beta_2)'(Y - X_1\beta_1 - X_2\beta_2) \\ &= (Y - X_2\beta_2)'M_{X_1}(Y - X_2\beta_2) + (\beta_1 - \hat{\beta}_1)'X_1'X_1(\beta_1 - \hat{\beta}_1) \\ &= Y'M_X Y + (\beta_2 - \hat{\beta}_2)'X_2'M_{X_1}X_2(\beta_2 - \hat{\beta}_2) + (\beta_1 - \hat{\beta}_1)'X_1'X_1(\beta_1 - \hat{\beta}_1) \end{aligned} \quad (2.89)$$

where  $\hat{\beta}_1 = (X_1'X_1)^{-1}X_1'Y$  and  $\hat{\beta}_2 = (X_2'M_{X_1}X_2)^{-1}X_2'M_{X_1}Y$ .

From Anderson (2003, ch.14):

$$(A_{22} - A_{21}A_{11}^{-1}A_{12})^{-1} = A_{22}^{-1}A_{21}(A_{11} - A_{12}A_{22}^{-1}A_{21})^{-1}A_{12}A_{22}^{-1} + A_{22}^{-1} \quad (2.90)$$

#### Orthogonal complements

From Johansen (1988)

$$\beta_{\perp}(\alpha'_{\perp}\beta_{\perp})^{-1}\alpha'_{\perp} + \alpha(\beta'\alpha)^{-1}\beta' = I, \quad (2.91)$$

where  $\beta'\alpha$  has full rank. If we choose  $\alpha = \beta = A^{-1/2}\tilde{\alpha}$  with  $\alpha_{\perp} = \beta_{\perp} = A^{1/2}\tilde{\alpha}_{\perp}$  (so that  $\alpha'\alpha_{\perp} = \tilde{\alpha}'\tilde{\alpha}_{\perp} = 0$ ), where  $\tilde{\alpha}$  has full rank, we have:

$$A^{1/2}\tilde{\alpha}_{\perp}(\tilde{\alpha}'_{\perp}A\tilde{\alpha}_{\perp})^{-1}\tilde{\alpha}'_{\perp}A^{1/2} + A^{-1/2}\tilde{\alpha}(\tilde{\alpha}'A^{-1}\tilde{\alpha})^{-1}\tilde{\alpha}'A^{-1/2} = I, \quad (2.92)$$

Pre- and postmultiplying with  $A^{-1/2}$  yields:

$$\tilde{\alpha}_{\perp}(\tilde{\alpha}'_{\perp}A\tilde{\alpha}_{\perp})^{-1}\tilde{\alpha}'_{\perp} + A^{-1}\tilde{\alpha}(\tilde{\alpha}'A^{-1}\tilde{\alpha})^{-1}\tilde{\alpha}'A^{-1} = A^{-1}. \quad (2.93)$$

#### Derivation of an matrix equation

**Theorem:** Consider two matrices  $A$  ( $T \times m_1$ ) and  $B$  ( $T \times m_2$ ), where  $m_1 \leq T$  and  $m_2 \leq T$ . Suppose that  $A$  has full rank, so that  $(A'A)^{-1}$  exists. Then we can decompose the determinant  $|(A B)'(A B)|$  as follows:

$$|(A B)'(A B)| = |A'A||B'M_A B| \quad (2.94)$$

where  $M_A$  is the  $T \times T$  projection matrix defined as  $M_A = I - A(A'A)^{-1}A'$ .

**Proof:** First, note that

$$|(A B)'(A B)| = \left| \begin{pmatrix} A'A & A'B \\ B'A & B'B \end{pmatrix} \right|. \quad (2.95)$$

and

$$|A'A|^{-1} = |(A'A)^{-1}| = \left| \begin{pmatrix} (A'A)^{-1} & 0 \\ 0' & I_{m_2} \end{pmatrix} \right|. \quad (2.96)$$

where 0 is the  $m_1 \times m_2$  zero matrix, and  $I_{m_2}$  is the  $m_2 \times m_2$  identity matrix. If we multiply equations (2.95) and (2.96), we have:

$$\begin{aligned} |(A B)'(A B)||A'A|^{-1} &= \left| \begin{pmatrix} (A'A)^{-1} & 0 \\ 0' & I_{m_2} \end{pmatrix} \right| \left| \begin{pmatrix} A'A & A'B \\ B'A & B'B \end{pmatrix} \right| \\ &= \left| \begin{pmatrix} (A'A)^{-1} & 0 \\ 0' & I_{m_2} \end{pmatrix} \begin{pmatrix} A'A & A'B \\ B'A & B'B \end{pmatrix} \right| \\ &= \left| \begin{pmatrix} I_{m_1} & (A'A)^{-1}A'B \\ B'A & B'B \end{pmatrix} \right|, \end{aligned} \quad (2.97)$$

where  $I_{m_1}$  is the  $m_1 \times m_1$  identity matrix. Multiplying (2.97) by

$$1 = \left| \begin{pmatrix} I_{m_1} & 0 \\ -B'A & I_{m_2} \end{pmatrix} \right| \quad (2.98)$$

we have:

$$\begin{aligned} |(A B)'(A B)||A'A|^{-1} &= \left| \begin{pmatrix} I_{m_1} & 0 \\ -B'A & I_{m_2} \end{pmatrix} \right| \left| \begin{pmatrix} I_{m_1} & (A'A)^{-1}A'B \\ B'A & B'B \end{pmatrix} \right| \\ &= \left| \begin{pmatrix} I_{m_1} & (A'A)^{-1}A'B \\ 0' & B'B - B'A(A'A)^{-1}A'B \end{pmatrix} \right| \\ &= |B'B - B'A(A'A)^{-1}A'B|. \end{aligned} \quad (2.99)$$

Finally, multiplying (2.99) by  $|A'A|$  yields:

$$\begin{aligned} |(A B)'(A B)| &= |A'A| |B'B - B'A(A'A)^{-1}A'B| \\ &= |A'A| |B'M_A B|, \end{aligned} \quad (2.100)$$

where  $M_A = I - A(A'A)^{-1}A'$ . □

**Corollary:** If additionally  $B$  has full rank, then

$$|B'M_A B| = \frac{|A'M_B A| |B'B|}{|A'A|}, \quad (2.101)$$

where  $M_B$  is the  $T \times T$  projection matrix defined as  $M_B = I - B(B'B)^{-1}B'$ .

**Proof:** Note that if  $B$  has full rank, we can switch the matrices  $A$  and  $B$ . So, in that case we have:

$$|A'A| |B'M_A B| = |(A B)'(A B)| = |(B A)'(B A)| = |B'B| |A'M_B A|. \quad (2.102)$$

and the result follows immediately. □

## 2.B Matricvariate distributions

### 2.B.1 Inverted Wishart distribution

Let  $\Sigma$  be an  $n \times n$  random symmetric positive definite matrix.  $\Sigma$  has an inverted Wishart distribution if its density function is

$$p(\Sigma|Q, \nu) = f_{IW}^m(\Sigma|Q, \nu) = c|\Sigma|^{-\frac{1}{2}(\nu+n+1)}|Q|^{\nu/2} \exp\left[-\frac{1}{2}\text{tr}(Q\Sigma^{-1})\right], \text{ for } |\Sigma| > 0 \quad (2.103)$$

where  $\Sigma$  is a symmetric positive definite  $n \times n$  matrix and  $\nu \geq n - 1$ . The constant  $c$  is given by  $c^{-1} = 2^{\frac{1}{2}\nu n} \Gamma_n(\nu/2)$  where

$$\Gamma_n(x) = \pi^{n(n-1)/4} \prod_{i=1}^n \Gamma[x + (1-i)/2] \quad (2.104)$$

is the multivariate gamma function.

If  $\Sigma$  has the above inverted Wishart density,  $\Psi = \Sigma^{-1}$  has a Wishart distribution with scale  $Q^{-1}$  and degrees of freedom  $\nu$ . An algorithm to generate random draws from an inverted Wishart distribution is derived in Zellner, Bauwens, and Van Dijk (1988, pp.67-71).

Using (2.103) it follows directly that

$$\int |\Sigma|^{-\frac{1}{2}(\nu+n+1)} \exp\left[-\frac{1}{2}\text{tr} Q\Sigma^{-1}\right] d\Sigma \propto |Q|^{-\frac{1}{2}\nu} \quad (2.105)$$

We refer to this integration result as the inverted Wishart step.

### 2.B.2 Matric-variate t distribution

The  $p \times q$  random matrix  $T$  has a matric-variate  $t$  (Mt) distribution (see Zellner (1971) and Kleibergen and Van Dijk (1998)) with parameters  $P$ ,  $Q$ ,  $n$  if, and only if, its probability density function is:

$$p_{Mt}(T|P, Q, n) = k \frac{|P|^{q/2} |Q|^{(n-p)/2}}{|Q + T' P T|^{n/2}} \quad (2.106)$$

$$= k \frac{|P|^{-(n-q)/2} |Q|^{-p/2}}{|P^{-1} + T Q^{-1} T'|^{n/2}} \quad (2.107)$$

where the equality follows from the following equality:

$$\frac{|Q + T' P T|}{|P| |Q|} = |P^{-1} + T Q^{-1} T'| \quad (2.108)$$

and where the constant  $k$  is given by:

$$k = \frac{1}{\pi^{pq/2}} \frac{\prod_{i=1}^q \Gamma[(n-i+1)/2]}{\prod_{i=1}^q \Gamma[(n-p-i+1)/2]}, \quad (2.109)$$

and we have  $n > p + q - 1$  and  $P$  and  $Q$  are positive definite symmetric (PDS) matrices of size  $p \times p$  and  $q \times q$ , respectively.

Alternatively, the matric-variate t-distribution is often parameterized in terms of the degrees of freedom parameter  $\nu = n - p$ .

**Marginal and conditional matrix-variate  $t$  distributions:** For a random matrix  $T$  which has a matrix-variate  $t$  distribution, marginal and conditional distributions of partitions of  $T$  also have a matrix-variate  $t$  distribution.

First note that (2.106) does not contain a location parameter. A location parameter  $M$  can be introduced using

$$p_{Mt}(T|M, P, Q, n) = p_{Mt}(T - M|P, Q, n). \quad (2.110)$$

Next, consider the following partitioning of  $T$ , location parameter  $M$  and the scale parameters  $P$ :

$$T = \begin{pmatrix} T_1^{(p_1 \times q)} \\ T_2^{(p_2 \times q)} \end{pmatrix}, \quad M = \begin{pmatrix} M_1^{(p_1 \times q)} \\ M_2^{(p_2 \times q)} \end{pmatrix}, \quad P = \begin{pmatrix} P_{11}^{(p_1 \times p_1)} & P_{12}^{(p_1 \times p_2)} \\ P_{21}^{(p_2 \times p_1)} & P_{22}^{(p_2 \times p_2)} \end{pmatrix} \quad (2.111)$$

where  $p_1 + p_2 = p$  and  $A^{(a,b)}$  denotes that matrix  $A$  has dimensions  $a \times b$ .

Then the following conditional and marginal densities hold:

$$p(T_1 | T_2, M, P, Q) = p_{Mt}(T_1 | M_{T_1|T_2}, P_{11}, Q_{T_1|T_2}, n), \quad (2.112)$$

$$p(T_2 | M, P, Q) = p_{Mt}(T_2 | M_2, P_{22.1}, Q, n - p_1), \quad (2.113)$$

where  $M_{T_1|T_2} = M_1 - P_{11}^{-1}P_{12}(T_2 - M_2)$ ,  $Q_{T_1|T_2} = Q + (T_2 - M_2)'P_{22.1}(T_2 - M_2)$  and  $P_{22.1} = P_{22} - P_{21}P_{11}^{-1}P_{12}$ .

See Zellner (1971), appendix B.5 and Bauwens et al. (1999), appendix A.2.

## Chapter 3

# Bayesian Analysis of the PPP Puzzle Using an Unobserved Components Model

*Chapter 3 is based on the working paper Kleijn and Van Dijk (2001), Econometric  
Institute Report EI 2001-35.*

### 3.1 Introduction

In its simplest form Purchasing Power Parity (PPP) is based on the Law of One Price, which relates domestic and foreign price levels ( $P_t$  and  $P_t^*$ ) to the nominal exchange rate  $S_t$ , i.e.

$$P_t = P_t^* S_t.$$

A more realistic version relaxes this equality, allowing for temporary deviations. Equivalently, this may be formulated as mean reversion of the real exchange rate  $Q_t$ . Using lower case symbols to denote logs, we have that

$$q_t = p_t^* - p_t + s_t$$

should be stationary.

There exists a vast literature on testing the hypothesis of PPP, often in the form of establishing whether the real exchange rates are stationary or nonstationary. Numerous unit root tests and stationarity tests such as the KPSS test of Kwiatkowski et al. (1992) and its variants have been used. In many cases these tests failed to reject the unit root or rejected stationarity, casting doubt on the validity of one of the fundamental building blocks of international economic theory. The relatively short span of available data on floating exchange rates in combination with the high persistence in real exchange rates could explain these findings. It is a known problem that it is very hard to distinguish between data containing a unit root and highly persistent but stationary data on the basis of a short span of data.

Moreover, real exchange rates typically show huge volatility in the short to medium term and at the same time it has a very high persistence with a half-life of deviations from its mean of 3 to 5 years. These features have given rise to the so-called Purchasing Power Parity Puzzle, see Rogoff (1996). The high volatility could potentially be explained by sticky prices and real shocks, but the mean reversion in the data is too slow to be consistent with these arguments.

On the other hand, mean reversion has been found for long spans of data of 100 years or more. This has convinced some researchers that the data provides evidence that real exchange rates really are stationary. However, Engel (2000) claims using a simulation study that these results may be spurious. For the real exchange rate data he finds that the unit root tests have serious size distortions while simultaneously the stationarity tests display a lack of power. He suggests that instead of temporary deviations from a fixed long-term mean it is more plausible to assume that the equilibrium rate itself is slowly changing over time. He explains this by distinguishing between traded goods for which price adjustment is likely and non-traded goods which lack an identifiable mechanism for mean reversion.

In this chapter we analyze the PPP puzzle as follows. First, we implement the suggestion of Engel (2000) that the long term mean of the real exchange rate may exhibit some time variation, by using an unobserved components model for the real exchange rate behavior. A long term component captures the time variation in the mean while short term components describe temporary deviations. In Section 3.3 we motivate why we have

chosen this particular model by assessing the pros and cons of existing approaches in the literature.

For inferential purposes we make use of a simulation based Bayesian analysis. The Bayesian approach has the advantage that one may give the hypotheses of stationarity and non-stationarity equal prior probabilities. Our posterior simulator for computing posterior distributions is introduced in Section 3.4. A third feature of our approach is exact inference on functions of parameters of interest, in particular we measure the half-life of deviations of the real exchange rate from its mean. Cheung and Lai (2000) conclude that classical point estimates of the half-life have a very large imprecision. They construct confidence intervals to quantify the uncertainty. Our Bayesian approach allows us to calculate the entire posterior density of the half-lives. In this chapter we also find non-monotonic impulse responses, consistent with their findings. Earlier Bayesian work on PPP includes Schotman and Van Dijk (1991) who use an autoregressive model with a fixed mean. Our simulation based Bayesian inference is an extension of the work of Koop and Van Dijk (2000), but their focus is on testing for stationarity in an unobserved components model. Bayesian work related to nonstationarity, but not to PPP specifically, include Sims (1988), Sims and Uhlig (1991), Phillips (1991) and Zivot (1994), or see Bauwens et al. (1999) for a survey. West and Harrison (1997) give an extensive Bayesian treatment of unobserved components models, but they do not focus on nonstationarity issues.

In this chapter, we can make a distinction between model selection for the PPP hypothesis using posterior odds and the analysis of PPP using the concept of half-life, which is carried out in the flexible model. The price of this flexibility is that one has to develop a tailored posterior simulator to evaluate the posterior density of functions of the parameters of interest. Given the methods of de Jong and Shephard (1995) or similar work of Carter and Kohn (1994) and Frühwirth-Schnatter (1994), one can develop such posterior simulators.

The outline of the chapter is as follows: we start by describing our data set in Section 3.2, followed by a description and motivation of the modelling process in Section 3.3. The details of the full Bayesian analysis of the model are set out in Section 3.4. In Section 3.5 we report posterior results. The results include tests for a constant underlying mean of the real exchange rates, posterior densities for some parameters of interest, the impulse responses of deviations from the (time-varying) mean and their implied half-lives. Section 3.6 concludes with a discussion of the results and an indication of possible extensions for future research on this topic.

## 3.2 Data

The data we use in this chapter are obtained from the OECD Statistical Compendium. The real exchange rates are constructed from nominal exchange rates and consumer price indices. The series have a monthly frequency and cover the post-Bretton Woods period, that is, they run from 1973:01 to 1998:12, when the internal Euro rates became fixed. We investigate the German Mark with the US dollar as numeraire currency (DEM/USD) and the French Franc against the German Mark (FF/DEM). Figure 3.1 displays the real exchange rates.

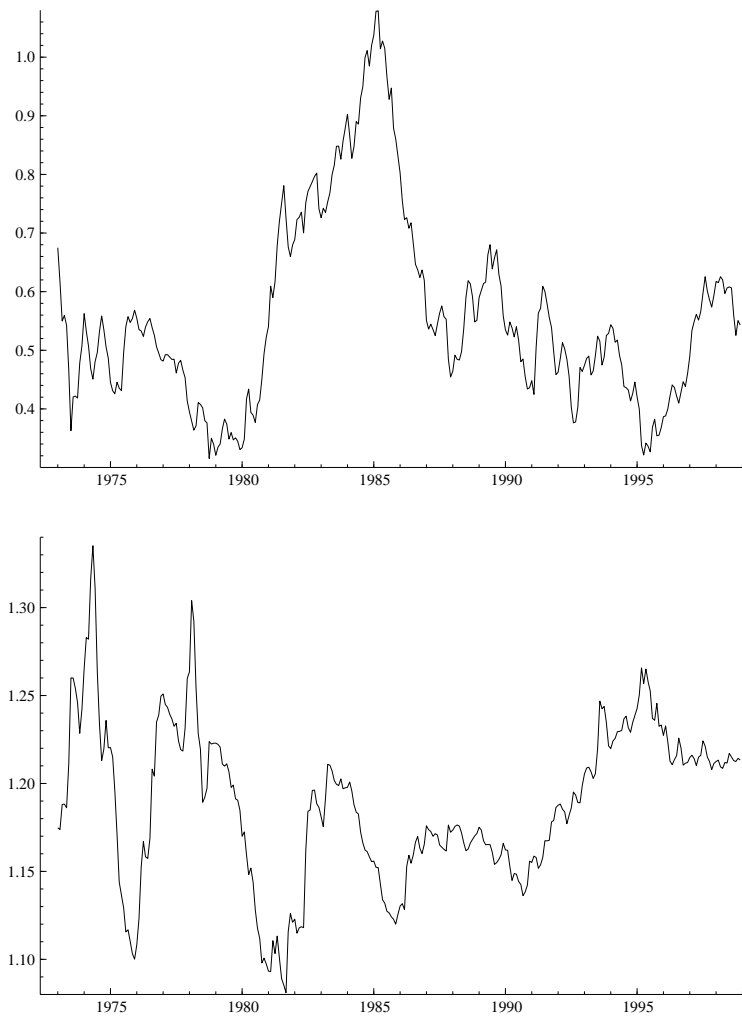


Figure 3.1: Real exchange rates (logs) of Germany with US as numeraire (DEM/USD) and France with Germany as numeraire (FF/DEM).

Germany and France are major trade partners, neighboring countries, and early European Union members. These factors lead us to expect beforehand that trading restrictions will be relatively minor and that the transportation costs are low. For this combination the real world circumstances are close to the assumptions underlying the theory of PPP.

### 3.3 Model Specification

In this section we motivate why we have chosen the particular model that we shall use. We do so by assessing the pros and cons of previously used approaches for testing the hypothesis of PPP and the ability of those models to describe the features of our data set adequately.



The most common approach is based on the ARIMA class of models. Unit root tests in many variants have been carried out in the context of this class. An important characteristic of the unit root tests is that they test stationarity of deviations from a fixed constant mean or from a rigid linear trend. The most common outcome for real exchange rate data is that the null hypothesis of a unit root cannot be rejected, possibly due to the high persistence of the deviations and the limited span of the data. In the ARIMA modelling paradigm the next logical step is to handle nonstationarity by analyzing the first differences of the series. As a result information on the level of the series is lost to some degree. An estimated ARIMA model can be decomposed in a random walk component and a stationary component by means of the Beveridge-Nelson decomposition, see Beveridge and Nelson (1981). The Beveridge-Nelson decomposition does indirectly what more structural unobserved components models can do explicitly and more transparently. An alternative for assessing whether the series is stationary or not is given by the KPSS test, see Kwiatkowski et al. (1992). They develop a test for the null hypothesis of the absence of a unit root. The setup they use is an unobserved components model, which is the sum of a random walk (possibly with drift) and a stationary component which captures the short run dynamics. The test consists of checking whether the random walk variance is close to zero such that the random walk reduces to a linear trend or a constant.

In classical testing the null and the alternative hypothesis are treated asymmetrically, implicitly favoring one outcome over the other a priori. In the Bayesian methodology equal prior probabilities can be assigned to two or more competing alternatives, in this case stationarity vs. nonstationarity. We shall treat the decomposition of the real exchange rate series into a nonstationary and a stationary component as a reasonable starting point for specifying our model. The simplest structural time series model is the so-called local level model (LLM). It consists of a random walk component with added i.i.d. normal errors. The drawback of this particular model is that under the null hypothesis the model reduces to i.i.d. normal deviations, which for our data set seems not to be very realistic. A straightforward way of improving on the behavior under  $H_0$  is to allow for more interesting dynamics in the stationary component.

Clark (1988) suggests a decomposition into a permanent random walk and transitory components. This approach has the same problem as the KPSS test. Testing whether the variance of the random walk is zero is problematic because estimation of the variance is imprecise. This phenomenon also causes the KPSS test to have low power.

The random walk component behaves erratically and is non-smooth. It picks up short and medium term dynamics. We consider this to be a second disadvantage of the simple LLM. We circumvent this problem by choosing a smooth I(2) trend instead of the ill-behaved random walk component. In the structural time series modelling literature a common means of imposing more smoothness is to replace the I(1) random walk with an integrated random walk of order I(2). It turns out that the variance parameter of the I(2) trend can be estimated with more precision, and hence we avoid issues associated with the low power problems of the KPSS test.

The spectral densities in Figure 3.2 show a large peak at a low frequency for both series. These frequencies are captured by the smooth trend component. The other peaks can be modelled adequately by two cycles. The main reason for choosing cycle components is

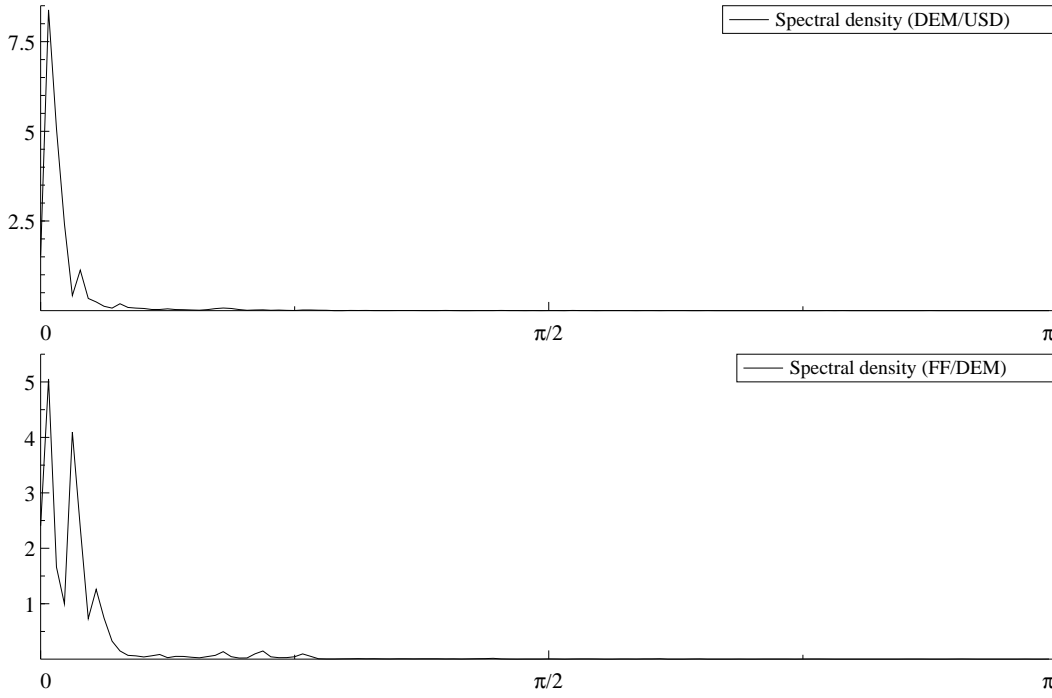


Figure 3.2: Sample periodogram for DEM/USD and FF/DEM.

that they are well-behaved in the sense that they do not tend to pick up the variation with a very low frequency that is associated with the trend component. We have experimented with stationary ARMA components for the short-run dynamics, but in our experience the ARMA parameters tend to move in the direction of the unit root in such a way that they start to interfere with the trend component. Although the cycles can be represented by an ARMA model, the implicit restrictions the cycle imposes on the ARMA reduced form prevent this pathological behaviour. Our preferred model thus contains a smooth trend and two cyclical components:

$$\begin{aligned}
 y_t &= \mu_t + \psi_{1,t} + \psi_{2,t} \\
 \mu_t &= \mu_{t-1} + \beta_{t-1} \\
 \beta_t &= \beta_{t-1} + \zeta_t \\
 \begin{pmatrix} \psi_1 \\ \psi_1^* \end{pmatrix}_t &= \rho_1 \begin{pmatrix} \cos \lambda_1 & \sin \lambda_1 \\ -\sin \lambda_1 & \cos \lambda_1 \end{pmatrix} \begin{pmatrix} \psi_1 \\ \psi_1^* \end{pmatrix}_{t-1} + \begin{pmatrix} \kappa_1 \\ \kappa_1^* \end{pmatrix}_t \\
 \begin{pmatrix} \psi_2 \\ \psi_2^* \end{pmatrix}_t &= \rho_2 \begin{pmatrix} \cos \lambda_2 & \sin \lambda_2 \\ -\sin \lambda_2 & \cos \lambda_2 \end{pmatrix} \begin{pmatrix} \psi_2 \\ \psi_2^* \end{pmatrix}_{t-1} + \begin{pmatrix} \kappa_2 \\ \kappa_2^* \end{pmatrix}_t
 \end{aligned} \tag{3.1}$$

with  $\zeta_t \sim N(0, \sigma_\zeta^2)$ , and  $\begin{pmatrix} \kappa_i \\ \kappa_i^* \end{pmatrix}_t \sim N(0, (1 - \rho_i^2) \sigma_{\psi_i}^2 I_2)$ . The parameters  $\lambda_1$  and  $\lambda_2$  are the frequencies and  $\rho_1$  and  $\rho_2$  are the damping factors of the damped stochastic sinusoids. The local mean of the series is provided by  $\mu_t$ . In the PPP analysis we shall test for  $\sigma_\zeta^2 = 0$ .

In a Bayesian analysis the initial conditions play a crucial role for correcting the pathological behaviour of the likelihood near the unit root, see e.g. Schotman and Van Dijk (1991) on initial conditions in the context of an AR(1) model. Hence, we assume that the initial values come from the unconditional distribution as recommended in e.g. Harvey (1989). Since  $\beta_t$  and  $\mu_t$  are non-stationary we assume that  $\beta_1$  and  $\mu_1$  have a diffuse distribution. The cycle components are initialized by  $\begin{pmatrix} \psi_i \\ \psi_i^* \end{pmatrix}_{t=1} \sim N(0, \sigma_{\psi_i}^2 I_2)$ . Note that the model can be represented as the sum of one ARIMA(0,2,0) and two ARMA(2,2) components, which results in a complicated ARIMA model when written in single-equation form.

### 3.4 Posterior simulator

We are primarily interested in conducting posterior inference on (functions of) the model parameters. For that purpose we need the posterior density  $p(\theta|Data)$  of the parameters  $\theta = (\sigma^2, \rho, \lambda)$  with  $\sigma^2 = (\sigma_\zeta^2, \sigma_{\psi_1}^2, \sigma_{\psi_2}^2)$ ,  $\rho = (\rho_1, \rho_2)$  and  $\lambda = (\lambda_1, \lambda_2)$ .

The posterior density is proportional to likelihood multiplied by prior. The likelihood for these unobserved components models involves an integral over the unobserved states  $L(y|\theta) = \int L(y, \alpha|\theta) d\alpha$ , where  $\alpha_t = (\mu_t, \beta_t, \psi_{1,t}, \psi_{1,t}^*, \psi_{2,t}, \psi_{2,t}^*)$  and quantities without a subscript  $t$  represent the vector containing all the elements. It can be calculated in terms of the prediction error decomposition which can be obtained from the Kalman filter.

Unfortunately, calculating posterior expectations analytically or sampling from the posterior distribution are complicated by the integral over the unobserved states that enter the posterior density through the likelihood. It is possible to circumvent these problems by recognizing that also the posterior occurs as a marginal distribution. Direct Bayesian analysis of  $p(\theta|Data)$  is analytically untractable, but if we extend the parameter space, the resulting model is easier to deal with. This technique is known as data augmentation, see Tanner and Wong (1987). We then analyze the extended model and in the end the results are marginalized with respect to the additional parameters in order to translate results back to the original model, i.e.

$$p(\theta|Data) = \int p(\theta, \alpha|Data) d\alpha.$$

In order to let the data dominate the prior in the posterior, we assume the standard noninformative priors for the parameters, that is

$$\begin{aligned} p(\sigma_\zeta^2) &\propto \sigma_\zeta^{-2} \\ p(\sigma_{\psi_i}^2) &\propto \sigma_{\psi_i}^{-2} \\ p(\rho_1, \rho_2) &\propto 1 \quad \text{for } \rho_1, \rho_2 \in [0, 1] \\ p(\lambda_1, \lambda_2) &\propto 1 \quad \text{for } \lambda_1, \lambda_2 \in [2\pi/T, \pi] \text{ and } \lambda_1 < \lambda_2 \end{aligned}$$

When we are calculating Posterior Odds ratios in the following section we shall replace some of the noninformative priors by weakly informative natural conjugate priors because non-informative priors are known to distort the outcome of the Posterior Odds analysis.

In the remainder of this section we explain our simulation algorithm which is the basis of our posterior inference. For a recent general survey of the use of simulation methods in Bayesian analysis, see Geweke (1999). For  $p(\theta, \alpha | \text{Data})$  we use a Gibbs sampler, see e.g. Gelfand and Smith (1990) or Casella and George (1992). One iteration of the Gibbs sampler for the joint distribution of several random variables consists of going through the sequence of drawing each of the random variables conditional on the most recently obtained value of the remaining variables. The algorithm consists of drawing from the following conditional posterior densities:

1.  $p(\alpha | \sigma^2, \rho, \lambda, \text{Data})$
2.  $p(\sigma^2, \rho | \alpha, \lambda, \text{Data}) = p(\sigma_\xi^2 | \beta) p(\rho_1 | \lambda_1, \psi_1, \psi_1^*) p(\sigma_{\psi_1}^2 | \rho_1, \lambda_1, \psi_1, \psi_1^*) \times$   
 $p(\rho_2 | \lambda_2, \psi_2, \psi_2^*) p(\sigma_{\psi_2}^2 | \rho_2, \lambda_2, \psi_2, \psi_2^*)$
3.  $p(\lambda_1 | \alpha, \sigma^2, \rho, \lambda_2, \text{Data}) = p(\lambda_1 | \sigma_{\psi_1}^2, \rho_1, \lambda_2, \psi_1, \psi_1^*)$
4.  $p(\lambda_2 | \alpha, \sigma^2, \rho, \lambda_1, \text{Data}) = p(\lambda_2 | \sigma_{\psi_2}^2, \rho_2, \lambda_1, \psi_2, \psi_2^*)$

The algorithm consists of four blocks of parameters and it can straightforwardly be extended to more than two cyclical components. Sampling from the density  $p(\alpha | \sigma^2, \rho, \lambda, \text{Data})$  is known as the simulation smoother, see de Jong and Shephard (1995) for more details on this method. Generating a draw from  $p(\sigma_\xi^2 | \beta)$  is straightforward. The situation can be described by the partial model  $\Delta\beta_t \sim N(0, \sigma_\xi^2)$ , which amounts to the textbook situation of a Gaussian zero mean, unknown variance problem. Having specified a natural conjugate prior, the posterior density of  $\sigma_\xi^2 | \beta$  has an inverted gamma distribution.

If it is hard to sample directly from the conditional density, but sampling from an approximating density is possible, a Metropolis-Hastings (M-H) step can be embedded in the Gibbs sampler. See e.g. Chib and Greenberg (1995) for an intuitive treatment of the M-H algorithm. Sampling  $\rho_i, \sigma_{\psi_i}^2 | \lambda_i, \psi_i, \psi_i^*$  is done using a M-H step with a candidate for  $\rho_i | \lambda_i, \psi_i, \psi_i^*$  that resembles the actual density closely and an exact draw from  $\sigma_{\psi_i}^2 | \rho_i, \lambda_i, \psi_i, \psi_i^*$ . For  $\lambda_1 | \rho_1, \sigma_{\psi_1}^2, \lambda_2, \psi_1, \psi_1^*$  and  $\lambda_2 | \rho_2, \sigma_{\psi_2}^2, \lambda_1, \psi_2, \psi_2^*$  we use a M-H step based on a carefully constructed Student-t candidate whose Taylor expansion around its mode matches the target density up to order 6. In the Appendix we provide more details on the steps of the algorithm which involve sampling the parameters in the cyclical components.

We have set up the simulation as follows. We ran the Gibbs sampler 255,000 times discarding the first 5,000 as a burn-in period. Of the remaining 250,000 we only used every 25th draw in order to eliminate most of the correlation between successive values. All posterior results have been calculated on the basis of the resulting 10,000 values of the parameters  $\theta$  and the states  $\alpha$ . In each 25th iteration also the full impulse response function for that parameter realization is calculated. On average the algorithm runs at more than 100 iterations/second on an AMD 1300MHz PC.<sup>1</sup>

---

<sup>1</sup>All computations were carried out using Ox, see Doornik (1999) and the SsfPack package (Koopman et al. (1998)).

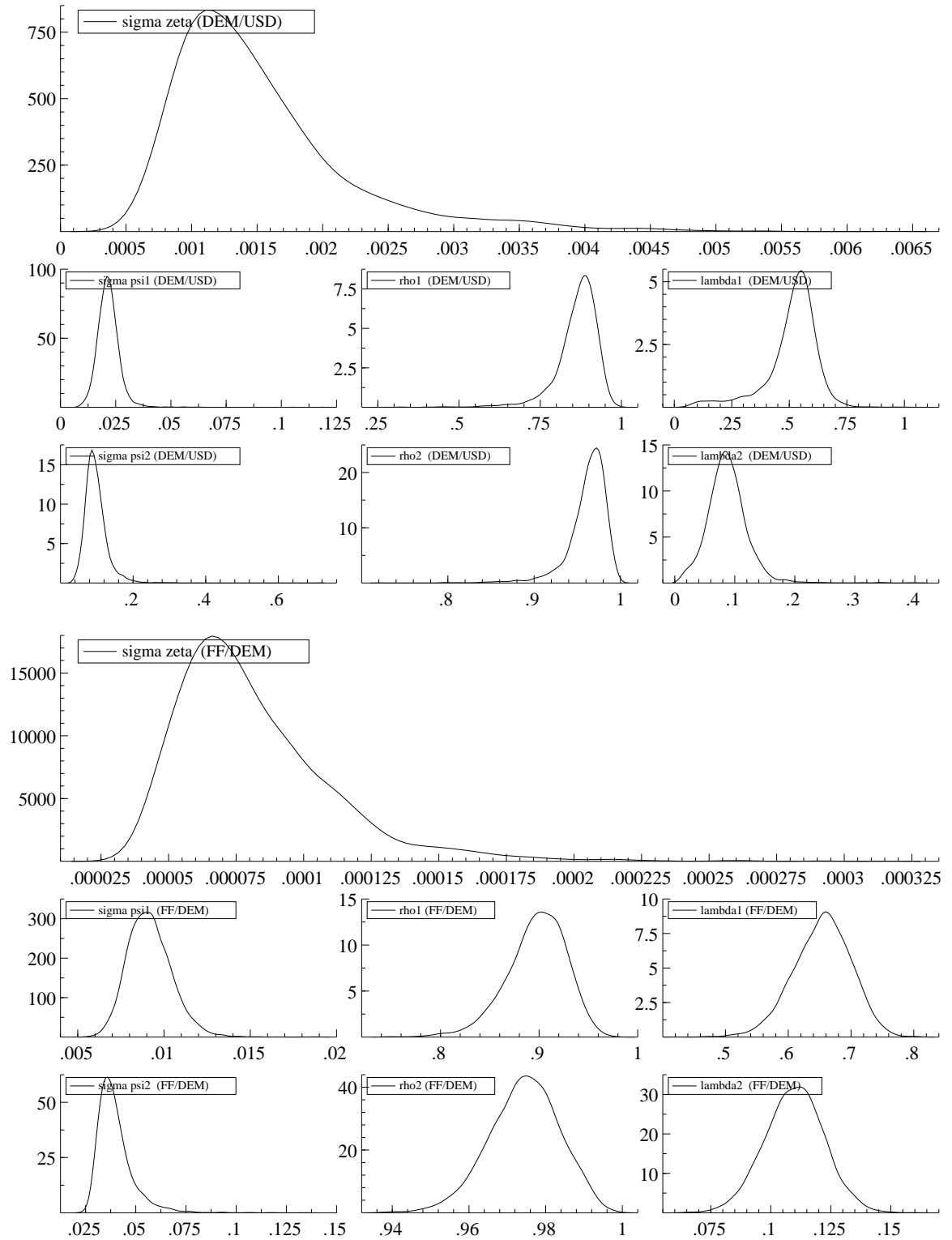


Figure 3.3: Posterior densities of model parameters for DEM/USD and FF/USD.

## 3.5 Posterior Results

### 3.5.1 Posterior distribution of the model parameters

The posterior densities of the parameters are graphed in Figure 3.3. Table 3.1 lists some summary statistics of the posterior distributions of the model parameters, namely means, standard deviations and medians. The most striking difference between the DEM/USD data and the FF/DEM data is the variance of the trend component. We shall come back to the implications of this in Subsection 3.5.3.

Table 3.1: Posterior results.

	DEM/USD		FF/DEM	
	mean	median	mean	median
$\sigma_\zeta (\times 10^{-3})$	1.52 (0.694)	1.355	0.0816 (0.0296)	0.0751
$\sigma_{\psi_1} (\times 10^{-3})$	21.7 (5.31)	21.4	9.18 (1.28)	9.09
$\rho_1$	0.863 (0.0666)	0.876	0.897 (0.0302)	0.900
$\lambda_1$	0.512 (0.114)	0.535	0.654 (0.0458)	0.656
$\sigma_{\psi_2} (\times 10^{-3})$	97.5 (31.8)	92.5	39.8 (9.07)	37.9
$\rho_2$	0.962 (0.0234)	0.966	0.974 (0.00919)	0.974
$\lambda_2$	0.0878 (0.0342)	0.0859	0.110 (0.0123)	0.110

*Note: The reported values are posterior means and medians obtained from the posterior simulator. Posterior standard deviations are given in parentheses.*

The medians of the periods of the cycles are 12.3 and 71.6 months, respectively, for the DEM/USD real exchange rate and we find 9.6 and 57.1 months for the FF/DEM data. The period of the shorter cycle may be related to a seasonal or calendar effect. The longer cycle may have a connection to the business cycle. However, including the cycle components serves merely as a means of providing the model with the required flexibility to capture any relevant short-run dynamics. Attaching a structural or economic interpretation to the individual cycles is not our primary focus.

### 3.5.2 Testing the sharp null hypothesis of stationarity

Posterior Odds ratios are the preferred Bayesian approach for testing a sharp null against a composite alternative. In this section we report the closely related Bayes Factors for the hypothesis  $\sigma_\eta^2 = 0$  which is a necessary condition for PPP. There are complications for interpreting Bayes Factors for a restriction if an improper prior is specified on the

parameters involved in the restriction of interest. Hence, we use weakly informative but proper inverted gamma priors for the variance parameters.

Chib's method for calculating Bayes factors, see Chib (1995), uses the output of a Gibbs sampler efficiently. Unfortunately, is not easily applicable because the constants of integration are not known for all conditional densities. Also the method by Chib and Jeliaskov (2001) has some drawbacks in this case, which we describe in more detail in Chapter 4.

Therefore, the marginal posteriors are computed using the Laplace approximation. The posterior distribution is approximated by a multivariate normal which has the posterior mode as its mean and its covariance matrix is minus the inverse of the Hessian at the posterior mode, see Kass et al. (1990). We transformed each parameter such that its support becomes the entire real line. As a side effect of this, the skewness of parameters is reduced, especially of the variance parameters. This helps making the approximation by a normal distribution more precise.

Table 3.2: More posterior results.

	DEM/USD	FF/DEM
BF	0.000357	0.000553
Half-life (mean)	13.6	14.8
Half-life (median)	12.0	11.0

*BF: the Bayes Factor for testing  $\sigma_\zeta^2 = 0$ . Half-life: mean and median of the posterior distribution of the half-life of the stationary components.*

The Bayes factors in Table 3.2 provide very strong evidence against  $\sigma_\zeta^2 = 0$ . The approximation error introduced by the Laplace expansion is unlikely to change this conclusion substantially.

### 3.5.3 Implied features of the model

An important byproduct of the Gibbs sampler is a sample of the smoothed components, for which the sample average over all draws at each point in time gives the Bayesian equivalent of the state smoother. The most important difference however is that parameter uncertainty is accounted for. The graphs of the expected value of the unobserved components are shown in Figures 3.4 and 3.5. Note that although the variation in the DEM/USD real exchange rates is much larger than for the FF/DEM data, the stationary component of both series have comparable variance. For the DEM/USD data more of the variation in the time series has been absorbed by the trend.

In the PPP literature there is a growing interest in the half-life of deviations from PPP. The half-life is defined as the number of years before the effect of a shock becomes permanently less than half the size of the original shock and therefore has a natural connection with impulse response functions. It is an important summary statistic of the full impulse response function. The half-life is often found to be between three and five years, see e.g. Cheung and Lai (2000) and the references cited therein. Note that this finding is based on models that impose PPP in a rigid way. In our model a fixed mean

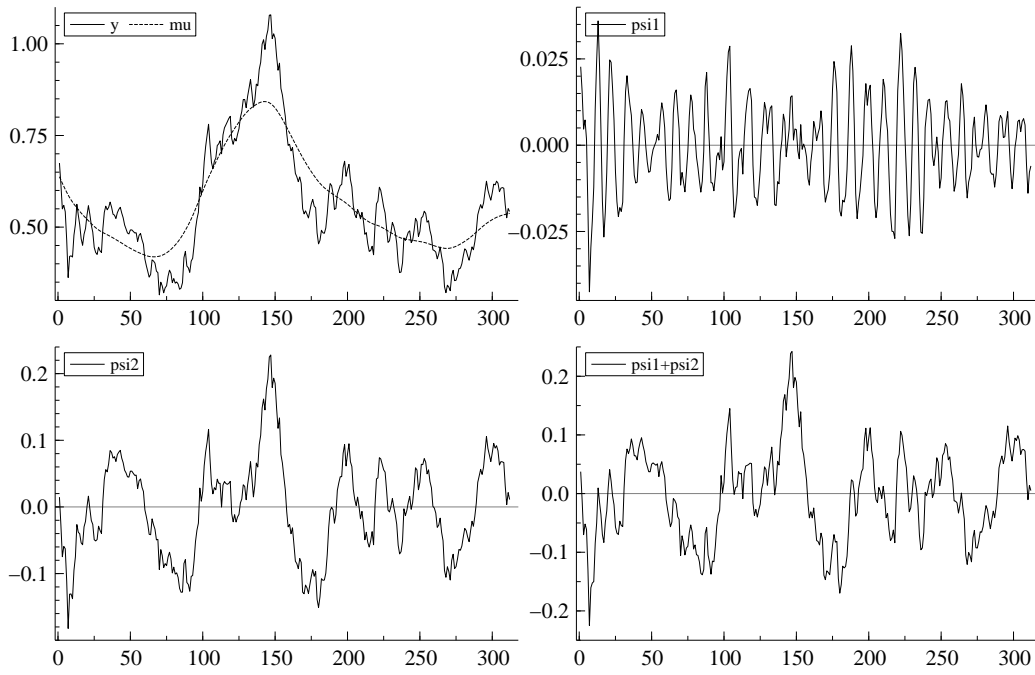


Figure 3.4: Posterior means of  $\mu_t$ ,  $\psi_{1,t}$  and  $\psi_{2,t}$  for DEM/USD.

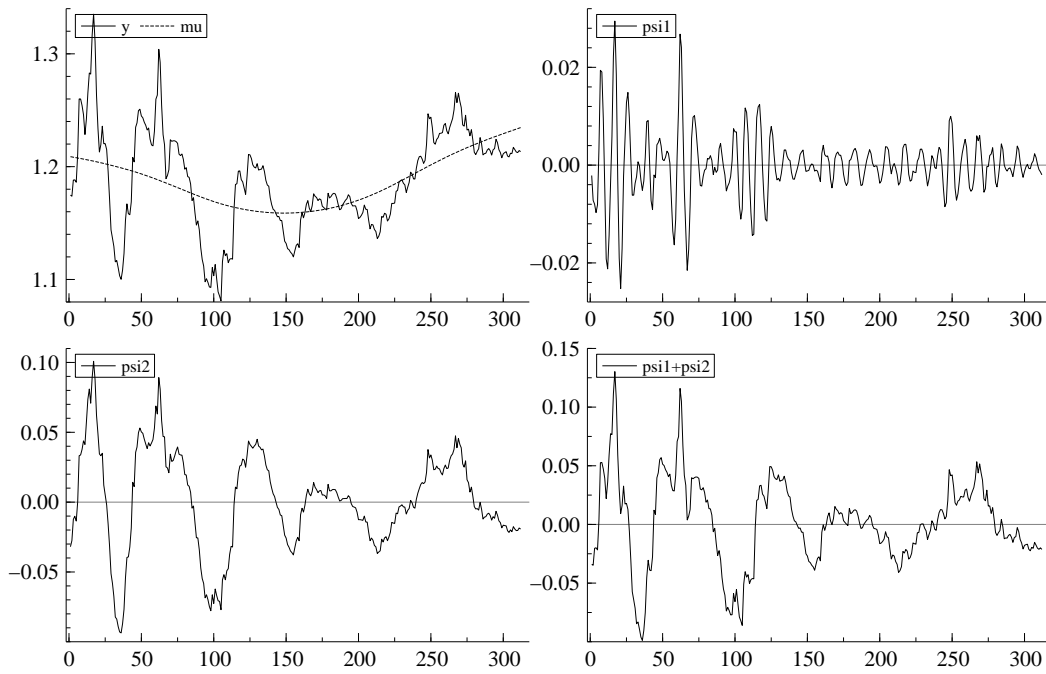


Figure 3.5: Posterior means of  $\mu_t$ ,  $\psi_{1,t}$  and  $\psi_{2,t}$  for FF/DEM.



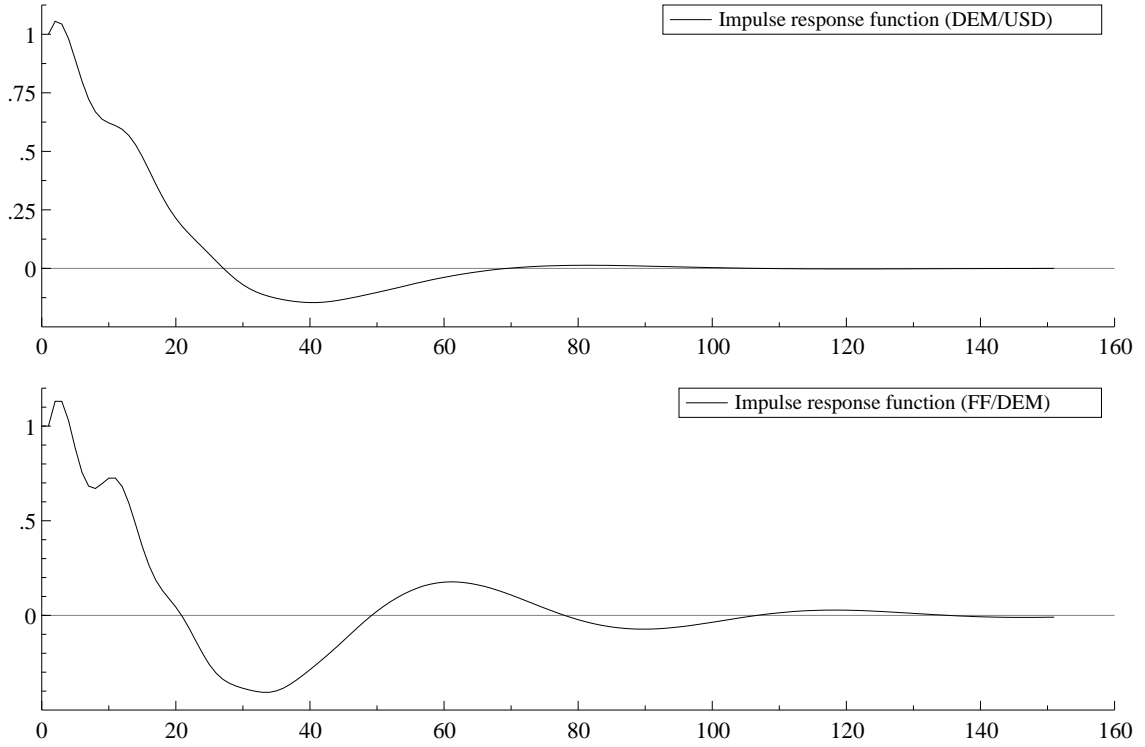


Figure 3.6: Posterior mean of the impulse response function of  $\psi_1 + \psi_2$  for DEM/USD and FF/DEM. The x-axis represents the horizon in months.

is not imposed. In our model, we first calculate the impulse response function of the deviations from the (time-varying) mean by means of the infinite moving average or Wold representation, see e.g. Hamilton (1994),

$$\Psi_i = ZT^{i-1}K, \quad (3.2)$$

where  $Z$  is the loading vector  $(1 \ 0 \ 1 \ 0)$  in the measurement equation,  $T$  is the block diagonal state transition matrix

$$\rho_1 \begin{pmatrix} \cos \lambda_1 & \sin \lambda_1 \\ -\sin \lambda_1 & \cos \lambda_1 \end{pmatrix} \oplus \rho_2 \begin{pmatrix} \cos \lambda_2 & \sin \lambda_2 \\ -\sin \lambda_2 & \cos \lambda_2 \end{pmatrix}, \quad (3.3)$$

and  $K$  is the  $4 \times 1$  steady-state Kalman gain which is straightforwardly obtained from the Kalman filter.

A unit shock is then assumed for  $t = 0$ . It is important to note that it matters which error term in our model is subjected to a shock. The value of the frequency parameter and the damping factor of that cycle influence how the effect of the shock will evolve over time. An unambiguous way of attributing the initial shock to the different components is to do so according to the Kalman gain. This provides an intuitive explanation of (3.2). Given the impulse response function, the half-life is trivially the time it takes for the impulse response to fall permanently below a half.

Different possible parameter configurations result in a variety of impulse responses. The Bayesian methodology enables us to integrate out the parameter uncertainty,

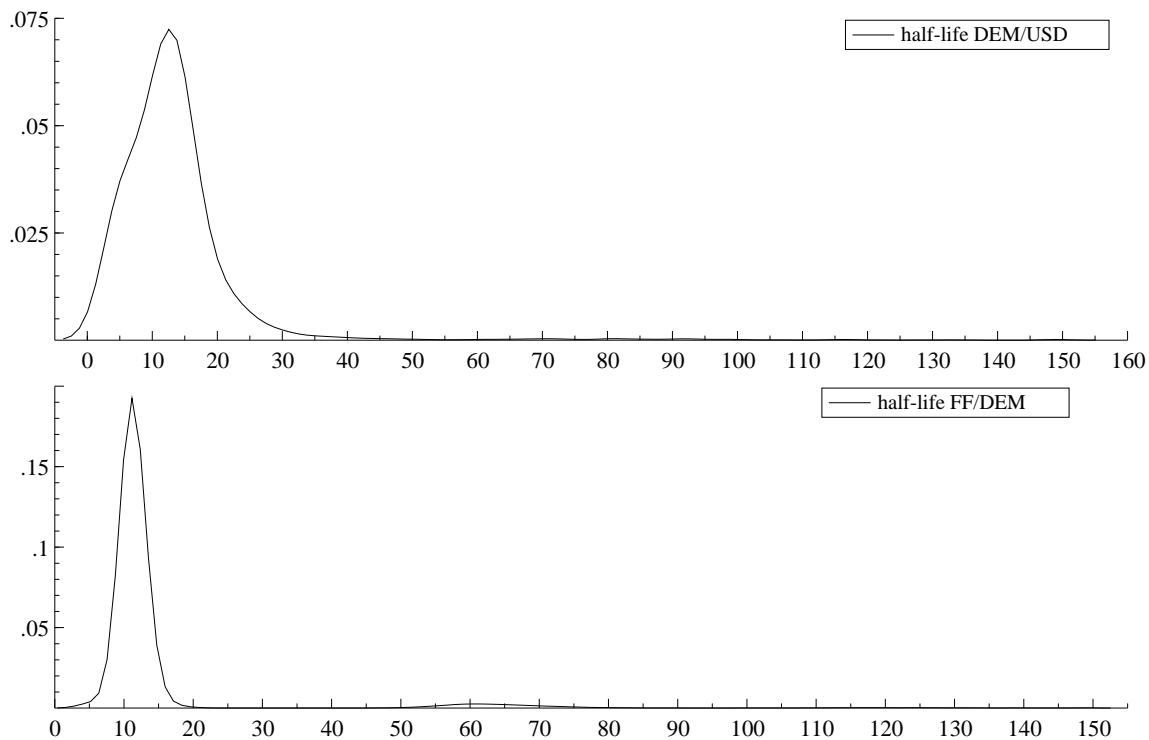


Figure 3.7: Posterior density of the half-life (in months) for DEM/USD and FF/DEM.

weighting each possible impulse response and its implied half-life according to the posterior probability of the particular outcome. Note that the posterior expectation of the half-life is not the same as the half-life of the expected impulse response, because the half-life is a highly nonlinear function of the impulse response.

In Figure 3.7 we see that the posterior density of the half-life even almost entirely excludes half-lives of 3 to 5 years. The half-lives we find are more likely to be attributable to sticky prices and other rigidities than the earlier findings. Including a flexible mean is the main reason that we find half-lives around one year for both real exchange rates. This is much shorter than the half-lives reported in previous studies which are in the range of 3 to 5 years. Surprisingly, we do not find substantial differences between the half-lives for the DEM/USD data and the FF/DEM data in the mean or median. From the densities of the half-lives, we find that for DEM/USD the distribution is much more dispersed than for FF/DEM, indicating less uncertainty for the intra-European combination of countries.

Note that we cannot use the full model to calculate half-lives. In that case the impulse responses are related to shocks to the entire system instead of shocks to the deviations from the mean. Part of the shock will be attributed to the trend component with the effect that the slope is changed. Thus, the initial shock will have a permanent and even increasing effect over time. The impulse response function will therefore diverge and no finite half-life exists in that case.

## 3.6 Final remarks

While the debate on whether long-run PPP holds or not will continue, we employ a pragmatic approach using an unobserved components model in which the importance of the mean reverting component relative to the non mean reverting component is allowed to vary. We provide some evidence for short run PPP for the recent period of floating exchange rates between the end of the Bretton Woods system and the introduction of the euro. After correcting for the long-term variation, the half-lives drop to reasonable levels of about one year. It seems that replacing the fixed-mean assumption of PPP by a more flexible slowly varying mean is definitely helpful in this respect. Our Bayesian framework can fully quantify the uncertainty of the half-life estimates that Cheung and Lai (2000) report by providing the entire posterior density of the half-lives. We find the same kind of non-monotonicity in the impulse response function as they do. A shock is initially amplified before it starts to die out.

Of course, the approach taken in this chapter is pragmatic in nature but at least it suggests a direction in which an explanation for the PPP puzzle may be found. It still remains a challenging task to explain and model the mechanisms that determine the time-varying behavior of the real exchange rates. Non-traded goods might not be the only factors responsible for the time-variation of the equilibrium rate. A viable other possibility would be to use a monetary exchange rate model that includes interest rates and money demand and supply. This is an interesting topic of further research. Furthermore, in the PPP literature panel models have aided a lot to compensate for the limited time span of the recent period of floating exchange rates. Hence, an extension of our model to panels of countries is a possible topic for future research.

## APPENDIX

### 3.A Sampling the cycle parameters

The cycle has three parameters, the variance  $\sigma^2$ , the damping factor  $\rho$  and the frequency parameter  $\lambda$ . We describe how to sample from their posterior distributions. The procedure is a Metropolis-Hastings (M-H) within Gibbs algorithm with carefully chosen candidate distributions. This algorithm for sampling of the cycle parameters can be used as a block of an encompassing Gibbs sampler when the cycle is a component of a larger model. For simplicity of the exposition we assume that a Jeffreys' prior is used for  $\sigma^2$  and for  $\rho$  and  $\lambda$  we specify uniform priors on  $[0, 1]$  and  $[0, \pi]$ , respectively. An inverted gamma prior on the variance is natural conjugate and can be handled analytically. Other priors can easily be implemented by taking account of them in the acceptance probabilities of the relevant Metropolis-Hastings (M-H) steps. Regarding the initial conditions we assume that  $\psi_1$  and  $\psi_1^*$  are initialized by a draw from the unconditional distribution of  $\psi = (\psi_1, \dots, \psi_T)$  and  $\psi^* = (\psi_1^*, \dots, \psi_T^*)$ . We introduce some other notation:  $Z' = (\psi', \psi^{*'})$ ,  $Z_{-1} = LZ$ , with  $L$  the lag operator, and  $C_\lambda = \begin{pmatrix} \cos \lambda & \sin \lambda \\ -\sin \lambda & \cos \lambda \end{pmatrix}$ .

The joint posterior density is obtained from combining the likelihood of a 2-dimensional restricted VAR(1) with the prior information and the initial conditions,

$$\begin{aligned} p(\rho, \lambda, \sigma_\kappa^2 | \psi, \psi^*) &\propto p(\rho, \lambda, \sigma_\kappa^2 | \psi, \psi^*) \\ &\propto L(\psi, \psi^* | \psi_0, \psi_0^*, \rho, \lambda, \sigma_\kappa^2) p(\psi_0, \psi_0^* | \rho, \sigma_\kappa^2) p(\rho, \lambda, \sigma_\kappa^2) \\ &\propto (\sigma_\kappa^2)^{-T} \exp\left(-\frac{1}{2\sigma_\kappa^2} \text{tr}((Z - \rho C Z_{-1})'(Z - \rho C Z_{-1}))\right) \times \\ &\quad (1 - \rho^2) \sigma_\kappa^{-2} \exp\left(-\frac{\psi_0^2 + \psi_0^{*2}}{2\sigma_\kappa^2}\right) \times \sigma_\kappa^{-2} \\ &= (1 - \rho^2) (\sigma_\kappa^2)^{-(T+2)} \exp\left(-\frac{1}{2\sigma_\kappa^2} [\text{tr}((Z - \rho C Z_{-1})'(Z - \rho C Z_{-1})) + \right. \\ &\quad \left. (1 - \rho^2)(\psi_0^2 + \psi_0^{*2})]\right) \end{aligned}$$

One can recognize the kernel of an inverted gamma distribution for  $\sigma^2$ , so the conditional posterior of  $\sigma^2$  is described by

$$\frac{\text{tr}((Z - \rho C Z_{-1})'(Z - \rho C Z_{-1})) + (1 - \rho^2)(\psi_0^2 + \psi_0^{*2})}{\sigma_\kappa^2} \Bigg| \rho, \lambda, \psi, \psi^* \sim \chi_{2(T+1)}^2.$$

Using an inverted gamma integration step, the joint marginal posterior density of  $\rho$  and  $\lambda$  is found to be

$$\begin{aligned} p(\rho, \lambda | \psi, \psi^*) &= \int p(\rho, \lambda, \sigma_\kappa^2 | \psi, \psi^*) d\sigma_\kappa^2 \\ &\propto (1 - \rho^2) [\text{tr}((Z - \rho C Z_{-1})'(Z - \rho C Z_{-1})) + (1 - \rho^2)(\psi_0^2 + \psi_0^{*2})]^{-(T+1)} \end{aligned}$$

We can rewrite the trace expression as a quadratic expression in  $\rho$

$$\begin{aligned} \text{tr}((Z - \rho CZ_{-1})'(Z - \rho CZ_{-1})) &= \text{tr}(Z'Z) - 2\rho \text{tr}(Z' CZ_{-1}) + \rho^2 \text{tr}(Z'_{-1} Z_{-1}) \\ &= \text{tr}(Z'_{-1} Z_{-1}) \left( \rho - \frac{\text{tr}(Z' CZ_{-1})}{\text{tr}(Z'_{-1} Z_{-1})} \right)^2 + \\ &\quad \left( \text{tr}(Z'Z) - \frac{\text{tr}(Z' CZ_{-1})^2}{\text{tr}(Z'_{-1} Z_{-1})} \right) \end{aligned}$$

The factor  $(1 - \rho^2)$  and the term  $(1 - \rho^2)(\psi_0^2 + \psi_0^{*2})$  represent the initial conditions of cycle. If enough observations are available, they tend to be dominated by the information in the data in the posterior density. Without these subexpressions one can recognize a Student-t kernel. Hence, it is sensible to use a (truncated) Student-t candidate in a M-H step for  $\rho|\lambda, \psi_1, \psi_2$ .

So,  $\rho^{cand}|\lambda, \psi, \psi^* \sim t(\mu, s^2, \nu)$ , i.e.  $p(\rho^{cand}) \propto \left(1 + \frac{(\rho^{cand} - \mu)^2}{\nu s^2}\right)^{-\frac{1}{2}(\nu+1)}$ , with

$$\begin{aligned} \mu &= \frac{\text{tr}(Z' CZ_{-1})}{\text{tr}(Z'_{-1} Z_{-1})} \\ s^2 &= \frac{1}{2T+1} \left[ \frac{\text{tr}(Z'Z)}{\text{tr}(Z'_{-1} Z_{-1})} - \frac{\text{tr}(Z' CZ_{-1})^2}{\text{tr}(Z'_{-1} Z_{-1})^2} \right] \\ \nu &= 2T+1 \end{aligned} \tag{3.4}$$

In a single M-H step we accept or reject the proposed values for  $\rho$  and  $\sigma^2|\rho$ .

Finally, the full conditional density of  $\lambda$  is given by

$$\begin{aligned} p(\lambda|\rho, \sigma_\kappa^2, \psi, \psi^*) &\propto p(\rho, \lambda, \sigma_\kappa^2|\psi, \psi^*) \\ &\propto \exp\left(\frac{\rho}{\sigma_\kappa^2} \text{tr}(Z' CZ_{-1})\right) \\ &= \exp\left(\frac{\rho}{\sigma_\kappa^2} [(\psi\psi'_{-1} + \psi^*\psi_{-1}^*) \cos \lambda + (\psi\psi_{-1}^* - \psi^*\psi'_{-1}) \sin \lambda]\right) \end{aligned}$$

We may combine the sine and cosine into one cosine with phase shift, and do a Taylor expansion as follows,

$$\begin{aligned} p(\lambda|\rho, \sigma_\kappa^2, \psi, \psi^*) &= \exp(A \cos \lambda + B \sin \lambda) \\ &= \exp(R \cos(\lambda - \varphi)) \\ &\propto 1 - \frac{1}{2}R(\lambda - \varphi)^2 + \frac{1}{24}R(3R+1)(\lambda - \varphi)^4 + O((\lambda - \varphi)^6) \\ &\approx \left(1 + \frac{(\lambda - \varphi)^2}{6}\right)^{-3R} \end{aligned}$$

where  $R^2 = A^2 + B^2$  and  $\varphi = \arctan(B/A)$ . The last expression is the kernel of a Student-t density with location  $\varphi$ , scale parameter  $\frac{6}{6R-1}$  and  $6R-1$  degrees of freedom. We use this Student-t distribution as a M-H candidate. The bounded support of  $\lambda$  introduces truncation, which can be accounted for either by using a truncated Student-t candidate or by including it in the acceptance probability of the independence Metropolis-Hastings step.



## Chapter 4

# Bayes Model Averaging of Cyclical Decompositions in Economic Time Series

*Chapter 4 is based on Kleijn and Van Dijk (2006).*

## 4.1 Introduction

Studying cyclical features of economic time series, frequency domain properties may provide a useful complement to time domain properties. The Fourier representation links time and frequency domain by showing that a linear combination of  $T$  orthogonal trigonometric terms can *perfectly fit* any given time series  $\{y_t\}_{t=1}^T$ . Implicitly it is assumed that the time series is exactly repeating itself infinitely far into the future. This regularity may be appropriate for some time series in e.g. geology, climatology, or astronomy, but economic time series tend to be less deterministic and more irregular. Moreover, the perfect fit prevents any useful statistical inferences to be made. As a descriptive procedure and data analysis tool the Fourier representation and the associated periodogram are helpful, but when e.g. forecasting is the objective, a model-based approach may be more appropriate.

In this chapter we propose a flexible model closely related to the Fourier representation

$$y_t = \mu + \sum_{c=1}^{\lfloor T/2 \rfloor} r_c \cos(\omega_c \pi t - \phi_c), \quad (4.1)$$

with  $\omega_c \equiv 2c/T$  and  $\phi_{\lfloor T/2 \rfloor} = 0$  when  $T$  is even, see e.g. Harvey (1993, Ch.6). Note that (4.1) could be put in linear form with orthogonal regressors by rewriting the terms  $r_c \cos(\omega_c \pi t - \phi_c)$  as  $a_c \cos \omega_c \pi t + b_c \sin \omega_c \pi t$ , where  $a_c$  and  $b_c$  are commonly referred to as the Fourier coefficients. When directly treated as a descriptive model for economic time series, (4.1) has  $\mu, r_1, \dots, r_{\lfloor T/2 \rfloor}$  and  $\phi_1, \dots, \phi_{\lfloor T/2 \rfloor}$  as its parameters, such that the number of parameters equals the number of observations.

We specify a model class that contains fewer than  $\lfloor T/2 \rfloor$  cycles and at the same time we introduce some randomness in the cycles. In this way the model does not constrain the data to be perfectly cyclical, while flexibility and exploitable structure are preserved. Our model class has the Fourier representation as its limiting case, thus ensuring that even for the messiest of time series at least one of the models will fit the data. Other model-based approaches include West (1995), Huerta and West (1999), and Harvey, Trimbur, and Van Dijk (2002). Huerta and West provide a method using Bayesian prior specification on parameters of a (high-order) AR model based on roots of the AR polynomial. They allow for uncertainty on the number of cycles, which are themselves directly associated with complex AR roots. In their approach all cycles are driven by a single innovation term, which may be overly restrictive as it introduces some complex structure in terms of cross-spectra. Harvey et al. describe an unobserved components structural time series model with a trend, an irregular term and a single cycle, but with the feature that each component has its own innovation term. They find prior information on the period of oscillation to be helpful in identifying the business cycle.

Summarizing, we introduce a model-based procedure based on an explicit unobserved components model which allows for stochastic cycles, as opposed to deterministic cycles. In particular we allow for an *unknown* number of cycles and use posterior odds analysis to decide on the number of cycles or, more appropriately, use its implied posterior model probabilities as weights in a mixture of models with different numbers of cyclical components to capture the inherent uncertainty on the number of cycles.



We illustrate our approach using simulated data. Next, several empirical applications show how this approach can be used as a tool for empirical work with wide applicability. For inferential purposes we make use of a Markov Chain Monte Carlo (MCMC) simulation based Bayesian analysis. The posterior simulator for our model class is based on data augmentation using a simulation smoother step within a Gibbs sampler. Our simulation based Bayesian inference is an extension of the work of Koop and Van Dijk (2000), but their focus is on testing for stationarity in an unobserved components model.

We present a convenient and useful reparametrization of the model which provides a structured way of specifying prior information. A particular contribution is the development of an efficient candidate generator for the model parameters of stochastic cycle components. It is specifically tailored for the Metropolis-Hastings step in the sampler for this model. The simulation based approach allows us to do inference directly on interesting functions of parameters by computing their full posterior distributions.

The outline of the chapter is as follows: In section 4.2 we start by introducing the model class that we are using throughout this chapter. We also outline the Bayesian procedures for these models. The artificial data experiment that concludes section 4.2 illustrates the use of these methods in practice. In sections 4.3 and 4.4 the cyclical decomposition procedure is applied to empirical time series, viz. US Industrial Production and unemployment. We investigate some relevant properties of each of the series under consideration. The results include tests for the constancy of the underlying mean, posterior model probabilities, posterior densities of parameters but also of some other quantities of interest which can be expressed as (more complicated) functions of the parameters and data. In particular we report full forecast densities, implied amplitudes, and spectral density estimates, which are notoriously difficult to estimate directly from the data. Section 4.5 concludes with a discussion of the results.

## 4.2 Cyclical Decomposition Procedure

For time series analysis the frequency domain often provides valuable information on the dynamics of the series. Other parts of the analysis are more appropriately done in the time domain. Frequency and time domain properties are more often than not analyzed separately. In this chapter we aim to provide a model for the time domain that more explicitly incorporates frequency domain properties with the underlying motivation that one domain could complement the other. In particular, our model consists of a composition of cycles, a feature it shares with the familiar discrete Fourier decomposition. This structure provides a very flexible functional form. We aim to obtain a structure that is as flexible as possible, without the danger of overfitting. We carefully try to avoid the noise fitting that would occur when fully unrestricted estimation would be tried. By expressing the model in terms of a hierarchical structure and specifying a prior accordingly we avoid these intricacies, while at the same time it facilitates the elicitation of prior information.

### 4.2.1 Specification

For policy-makers it is often relevant to be able to directly target either the structural part or the cyclical part of a time series. This requires insight in what establishes the permanent and the transitory components by means of making a trend-cycle decomposition. One such approach is that by Harvey and Jaeger (1993), who use an orthogonal unobserved components model to decompose the time series into a trend, a single cycle and an irregular component. In contrast, Morley, Nelson, and Zivot (2003) propose to introduce correlations between the components. Including such correlations in a model with more than one cycle requires complicated restrictions on the correlation structure in order to maintain identifiability of the model parameters from the autocovariance function. Hence, we follow the Harvey-Jaeger approach with orthogonal components as it generalizes in a transparent way to multiple cycles. A single cycle may not be adequate to capture the dynamics of the cycle components, hence we allow for multiple cycles as we want to assess the merits of long, medium, short or seasonal cycles without restricting them beforehand. In this context, the business cycle is conventionally associated with frequencies corresponding to cycle lengths from 6 to 32 quarters.

The core of the model consists of a composition of several (unobserved) cyclical components. We use the same functional form as e.g. Harvey and Streibel (1998). The (restricted) two-dimensional VAR(1) model of (4.2c) ensures a regular cycle with a pronounced peak in the theoretical spectrum. These stochastic cycles become smoother when the damping factor gets closer to one, having a deterministic cycle as their limiting case. A simple AR(1) is the limiting case when the frequency parameter tends towards zero. The model also contains a trend component and a noise term.

Thus, our model consists of components meant to describe long-term movements, cyclical features and irregular movements of the data, and is given as

$$y_t = \mu_t + \sum_{c=1}^C \psi_{c,t} + \xi_t \quad (4.2a)$$

$$\mu_t = \mu_{t-1} + \delta + \eta_t \quad (4.2b)$$

$$\begin{pmatrix} \psi_{c,t} \\ \psi_{c,t}^* \end{pmatrix} = \rho_c \begin{pmatrix} \cos \omega_c \pi & \sin \omega_c \pi \\ -\sin \omega_c \pi & \cos \omega_c \pi \end{pmatrix} \begin{pmatrix} \psi_{c,t-1} \\ \psi_{c,t-1}^* \end{pmatrix} + \begin{pmatrix} \kappa_{c,t} \\ \kappa_{c,t}^* \end{pmatrix}, \quad \text{for } c = 1, \dots, C \quad (4.2c)$$

for which we assume Gaussian innovations  $\xi_t \sim N(0, \sigma_\xi^2)$ ,  $\eta_t \sim N(0, \frac{6}{T} \sigma_\mu^2)$ , and  $(\kappa_c, \kappa_c^*)_t \sim N(0, (1 - \rho_c^2) \sigma_c^2 I_2)$ , for  $c = 1, \dots, C$ .

The trend  $\mu_t$  provides the local mean of the series and is specified in (4.2b) as a random walk with drift rate  $\delta$ . It could also be replaced by a smoother trend specification. In our implementation, the drift rate is absorbed in the state vector. Note that  $\sigma_\mu^2$  here represents the expected sample variance in a sample of size  $T$  from the process  $\{\mu_t\}$ , which does exist whereas the unconditional variance does not. The model contains  $C$  damped stochastic sinusoids  $\psi_c$ . Each cycle has a frequency parameter  $\omega_c$  and a damping factor  $\rho_c$ . The closer the damping factor is to one, the smoother the cycle will be. The variance  $\sigma_c^2$  has the interpretation of the unconditional expected amplitude of the cycle. The (latent) amplitude of a stochastic cycle  $\psi_t$  at a specific time  $t$  is defined as  $\sqrt{\psi_t^2 + \psi_t^{*2}}$ . The (latent) phase of a stochastic cycle  $\psi_t$  at a given point in time  $t$  is given by  $\arctan \frac{\psi_t}{\psi_t^*}$ .

and hence the initial values  $\psi_0$  and  $\psi_0^*$  determine the initial value of the phase of the cycle. The irregular term  $\xi_t$  is specified as additive noise. For future use we define  $\mu = (\mu_1 \dots \mu_T)'$ ,  $\Psi_c = (\psi_{c1} \dots \psi_{cT})'$ ,  $\Psi_c^* = (\psi_{c1}^* \dots \psi_{cT}^*)'$ ,  $\alpha = (\mu \Psi_1 \dots \Psi_C \Psi_1^* \dots \Psi_C^*)$  and  $\theta = (\sigma_\mu^2 \sigma_\xi^2 \sigma_1^2 \dots \sigma_C^2 \omega_1 \dots \omega_C \rho_1 \dots \rho_C)'$ .

In a Bayesian analysis the initial conditions play a crucial role for correcting the pathological behaviour of the likelihood near the unit root, see e.g. Schotman and Van Dijk (1991) on initial conditions in the context of an AR(1) model or Zivot (1994) within an unobserved components model. We specify a bivariate normal distribution for the initial values of each cycle, with its mean and covariance matrix chosen to match the unconditional moments, i.e.  $(\psi_{c,1}, \psi_{c,1}^*) \sim N(0, \sigma_c^2 I_2)$ . We assume  $\mu_1$  to be fully diffuse, such that it acts as an intercept for the entire model.

The model essentially consists of the sum of a random walk, several ARMA(2,1) components and a white noise term, which results in a high-order ARIMA model with complicated parameter restrictions when represented in single-equation form.

### 4.2.2 Parameterization

It is our experience that in these kinds of unobserved components models the likelihood is often ill-behaved and suffers from (local) non-identification issues or multimodality. Maximum Likelihood and naive diffuse Bayesian analysis alike may easily lead to implausible results. Hence, it is of importance to understand the role of the model parameters and the structure of the unobserved components model and how they are connected. Prior elicitation and specification, as well as the performance of MCMC algorithms will benefit from dealing properly with these issues, see e.g. Koop and Van Dijk (2000). In a model with potentially many components it is particularly important that the components can be clearly distinguished and are not overlapping. If this is not the case, the components may spuriously pick up other frequencies and hence hamper identification of the individual components.

The unconditional variance of the trend component is not finite (unless  $\sigma_\mu^2 = 0$ ). We propose to compare the unconditional variance of the cycles with the expected sample variance of the random walk over the sample period. The latter number is finite and hence is a metric directly comparable to the variances of the cyclical components. This corresponds to the intuition that with a longer span of data one can identify cycles with larger periods. The expectation of the overall sample variance  $\sigma^2$  is divided between the trend term (fraction  $\nu$ ) and the remaining stationary components. This can be described as an hierarchical structure on the variances. We therefore introduce a transformation

from  $\theta$  to  $\tilde{\theta} \equiv (\sigma^2 \ \nu \ \lambda_1 \ \dots \ \lambda_C \ \omega_1 \ \dots \ \omega_C \ \rho_1 \ \dots \ \rho_C)'$  characterized by

$$\begin{aligned}\sigma_\mu^2 &= \nu \sigma^2 \\ \sigma_\psi^2 &\equiv \sigma_\xi^2 + \sum_{c=1}^C \sigma_c^2 = (1 - \nu) \sigma^2 \\ \sigma_c^2 &= \lambda_c \sigma_\psi^2 \\ \sigma_\xi^2 &= (1 - \sum_{c=1}^C \lambda_c) \sigma_\psi^2\end{aligned}$$

with Jacobian  $\left| \frac{\partial \theta}{\partial(\tilde{\theta})'} \right| = (1 - \nu)^C (\sigma^2)^{C+1}$ . We refer to the Appendix for details on the derivation. Thus, the Jacobian factor has a simple expression and allows us to operate in the transformed space in a relatively easy way.

### 4.2.3 Prior

This parametrization simplifies eliciting priors, both informative and noninformative. An example of a weakly informative prior (that is, relative with respect to the information in the likelihood) is obtained when we specify for the overall variance a diffuse prior on its log. The variance of the cyclical components is given as a fraction of the overall variance. Dirichlet priors are the standard choice for the fraction parameters. The prior that we use throughout the chapter is

$$\nu \sim B\left(\frac{1}{2}, \frac{1}{2}\right) \quad (4.3a)$$

$$(\lambda_1, \dots, \lambda_C) \sim \text{Dirichlet}(1, \dots, 1) \quad (4.3b)$$

$$(\omega_1, \dots, \omega_C) \sim U(0 \leq \omega_1 \leq \dots \leq \omega_C \leq 1) \quad (4.3c)$$

$$\rho_c | \omega_c \sim B(2/\omega_c, 1), \quad \text{for } c = 1, \dots, C \quad (4.3d)$$

Here the prior on  $\omega_c$  is uniform on a triangular subset to ensure identification of the cycles by ordering them according to frequency. This could alternatively be written as standard Dirichlet on the increments:  $(\omega_1, \omega_2 - \omega_1, \dots, \omega_C - \omega_{C-1}) \sim \text{Dirichlet}(1, \dots, 1)$ . The parameters of the Dirichlet distributions can also be made dependent on the number of cycles to ensure that the frequencies are centered closer around an equidistant grid when the number of cycles increases. In that case the prior could be specified as  $(\omega_1, \omega_2 - \omega_1, \dots, \omega_C - \omega_{C-1}) \sim \text{Dirichlet}(h_C, \dots, h_C)$ , with  $h_C$  a parameter that depends on the number of cycles in the model. The closer centering could for instance be accomplished by using  $h_C = C \frac{\lfloor T/2 \rfloor - 1}{\lfloor T/2 \rfloor - C}$ . For  $C \rightarrow \lfloor T/2 \rfloor$  the parameter  $h_C$  goes to infinity. This reflects that in the limiting case of  $\lfloor T/2 \rfloor$  cycles the distribution of the frequencies should degenerate to the fixed Fourier frequencies without any spread around them. This improves identification of the cycles. For small values of  $C$ ,  $h_C$  is approximately equal to  $C$ .

The period of the  $c$ -th cycle is  $2/\omega_c$ . In each observation period only a fraction  $\rho_c$  of the information on amplitude and phase of the  $c$ th cycle is carried forward. After a

complete cycle only a fraction  $\rho_c^{2/\omega_c}$  of the initial information remains. The value  $\rho_c^{2/\omega_c}$  can be thought of as a measure of smoothness of the cycle. In contrast to  $\rho_c$  it is comparable between cycles of different frequencies. The interpretation in terms of smoothness of the cycles is helpful for prior elicitation. We have chosen to specify a uniform prior on smoothness. This implies the mentioned beta distributions for the damping parameters  $\rho_c$ .

The prior we specified is improper. This impropriety is strictly due to  $p(\sigma^2)$ . This poses no problem for posterior odds analysis as this (nuisance) parameter occurs in all models throughout the entire model class with exactly the same interpretation and prior distribution. Hence, it cancels in the ratio of marginal likelihoods. It serves merely as a scale parameter and has similar orthogonality properties as the variance parameter in the linear regression case and therefore our procedure is scale invariant. Additionally, one can specify prior information on the number of cycles  $C$ . Straightforward examples are e.g. equal probabilities, (truncated) geometric or Poisson. This prior acts as a tuning parameter to influence the tendency of overfitting. This matter will be discussed in more detail in subsection 4.2.6.

#### 4.2.4 Posterior Simulator

We are primarily interested in conducting posterior inference on (functions of) the model parameters. For that purpose we need the posterior density  $p(\theta|Data)$  of the model parameters. Direct Bayesian analysis of  $p(\theta|Data)$  is analytically intractable, but we can apply the technique of data augmentation. We extend the parameter space by treating the unobserved states  $\alpha$  as extra parameters and the resulting joint posterior  $p(\theta, \alpha|Data)$  is much easier to simulate from.

We sample from the extended vector and by marginalizing with respect to the additional states we can easily translate results back to the original parameters, i.e.

$$p(\theta|Data) = \int p(\theta, \alpha|Data) d\alpha.$$

In the remainder of this section we explain our simulation algorithm which is the basis of our posterior inference. For  $p(\theta, \alpha|Data)$  we use a Gibbs sampler. One iteration of the Gibbs sampler for the joint distribution of several random variables consists of going through the sequence of drawing each of the random variables conditional on the most recently obtained value of the remaining variables. The algorithm consists of alternately drawing from the conditional densities  $p(\alpha|\theta, y)$  and  $p(\theta|\alpha, y)$ . The first step is implemented as a simulation smoother. We employ the version by de Jong and Shephard (1995). The implementation of the second step is a more intricate matter.

An inherent difficulty with data-augmentation is that the strong dependence of sampled model parameters on the state vector, and the other way around, introduces strong correlations in the Markov chain and this leads to relatively slow mixing. Therefore, it is essential to sample model parameters conditionally on the states as efficient as possible in the sense that it should not introduce much additional correlation. Hence, it is desirable to sample all model parameters in one single Gibbs step. Direct sampling of all model parameters simultaneously does not appear feasible

here, but we have developed an efficient, specifically tailored Metropolis-Hastings step which comes close. It is based on a candidate density which is both convenient for sampling and closely approximates the true conditional posterior density. We refer to the appendix for technical details.

### 4.2.5 Model Choice and Model Averaging

Now that we are able to analyze the model for a given number of cycles the question arises which number of cycles to use. Bayesian decision-theory provides a framework for answering this issue and we shall restrict our attention to the special cases of model choice and model averaging based on posterior odds ratios. We note that half the sample size is a natural upperbound on the number of cycles arising from the Fourier representation in (4.1). For our purposes imposing a maximum of 4 or 5 cycles seems reasonable as it saves computation time and otherwise we would merely be fitting noise with the extra cycles. We would like to note that our method is quite robust for overfitting because the models with more cycles than necessary are given low posterior model probability. The results in section 4.2.6 support this claim. An intuitive explanation for this phenomenon is that the parameter uncertainty in that case starts to outweigh the ‘certainty’ introduced by the better fit. If one assigns equal prior probabilities to all models in the class of models taken into consideration, the resulting ratio of the posterior probabilities of the models is a Bayes factor which can be calculated as the ratio of marginal likelihoods.

Several algorithms are available for calculating Bayes factors, either directly or through marginal likelihoods, many of which are unsuitable here because of multimodality of the posterior or because the integrating constants of the conditional distributions are unavailable. We have experimented with the harmonic mean estimator and with the method by Chib and Jeliaskov (2001). The harmonic mean estimator is obtained from the posterior sample by the likelihood values evaluated at the sampled  $\theta^{(i)}$  as

$$\frac{1}{\hat{m}(y)} = \frac{1}{n} \sum_{i=1}^n \frac{1}{L(y|\theta^{(i)})} \quad (4.4)$$

In practice we use the numerically more stable log-likelihoods  $LL_i = \log L(y|\theta^{(i)})$  and their minimum  $\underline{LL}$

$$\log \hat{m}(y) = -\log \frac{1}{n} \sum_{i=1}^n \frac{1}{L(y|\theta^{(i)})} = \underline{LL} - \log \frac{1}{n} \sum_{i=1}^n \exp(\underline{LL} - LL_i)$$

The Chib-Jeliaskov method calculates the marginal likelihood from the M-H output and an additional Markov chain from the identity

$$m(y) = \frac{L(y|\theta^*)\pi(\theta^*)}{\pi(\theta^*|y)}, \quad \text{for any } \theta^* \text{ in the support of } \theta$$

The likelihood and prior can be evaluated exactly. Let  $q(\theta|\theta^{(-1)}, \alpha, y) = q(\theta|\alpha, y)$  denote the candidate generating function of the Metropolis-Hastings step, where  $\theta^{(-1)}$  can be suppressed in our independence sampler. Also, let  $AP(\theta|\theta^{(-1)}, \alpha, y)$  denote the probability

of a move. Then, the posterior density can be calculated from the Metropolis-Hastings sampler as

$$p(\theta|y) = \frac{E_{\theta,\alpha} [AP(\theta^*|\theta, \alpha, y)q(\theta^*|\alpha, y)]}{E_{\alpha|\theta^*} [AP(\theta|\theta^*, \alpha, y)]}$$

in which  $E_{\theta,\alpha}$  represents the expectation with respect to  $p(\theta, \alpha|y)$  and  $E_{\alpha|\theta^*}$  refers to the expectation based on  $\pi(\alpha|\theta^*, y) \times q(\theta|\theta^*, \alpha, y)$ . The former expectation is computed as a sample mean based on a run of the posterior sampler from the previous section. The latter is based on an auxiliary run in which the required quantities are generated.

The results from both methods are qualitatively similar although there are some stability issues in both. The harmonic mean estimator is known to converge almost surely, but draws from the tails of the posterior distribution, i.e. draws with a low likelihood value, tend to be relatively influential in the resulting estimate. When monitoring the sampling process a pattern emerges of slow hyperbolic increase in the estimate which is subsequently offset by incidental downward jumps. The jumps become smaller as the number of iterations is increased, and we continue until the required accuracy is obtained. The results in this chapter are based on this estimator. The Chib-Jeliaskov approach suffers from the conditional posterior of the model parameters given the unobserved states being very tight and often very different from their marginal distribution. Hence, the factor  $q(\theta^*|\alpha, y)$  exhibits very high variability. If  $\theta^*$  is in the center of the conditional distribution, the value of  $q(\theta^*|\alpha, y)$  can be huge and when  $\theta^*$  is in the tails,  $q(\theta^*|\alpha, y)$  is virtually zero. This is not easily solved by increasing the number of iterations of the sampler. A partial solution would be to use a wider proposal distribution  $q(\theta^*|\alpha, y)$ . This comes at the cost of deteriorating the Markov chain as the draws from the tails would sparsely be accepted, thus lowering the acceptance rate. The numerical accuracy of alternative methods to evaluate the Bayes factor in state space models with a high-dimensional augmented parameter vector is of considerable interest but outside the scope of the present thesis.

## 4.2.6 Artificial Data Experiments

We illustrate the procedure with two artificial datasets. Each provides an important insight. The first experiment is based on simulated data with very smooth cycles. The second data set is based on less regular cycles.

### Artificial data with smooth cycles

We have simulated 3 series of 300 observations. The first series has 1 cyclical component with a cycle length of 60 periods. The second series additionally contains a cycle of 12 periods and the third data set adds an 8-period cycle. The series contain no trend component. The damping factors are chosen close, but not equal, to 1. The overall variance  $\sigma^2$  is equal to 1. The variance shares  $\lambda_c$  of the cycles are chosen such that the variance of the first cycle is twice as large as the variance of the second cycle and four times the variance of the third cycle. The irregular term receives a very small share of the variance. The generated series are shown in the first three panels of figure 4.1.

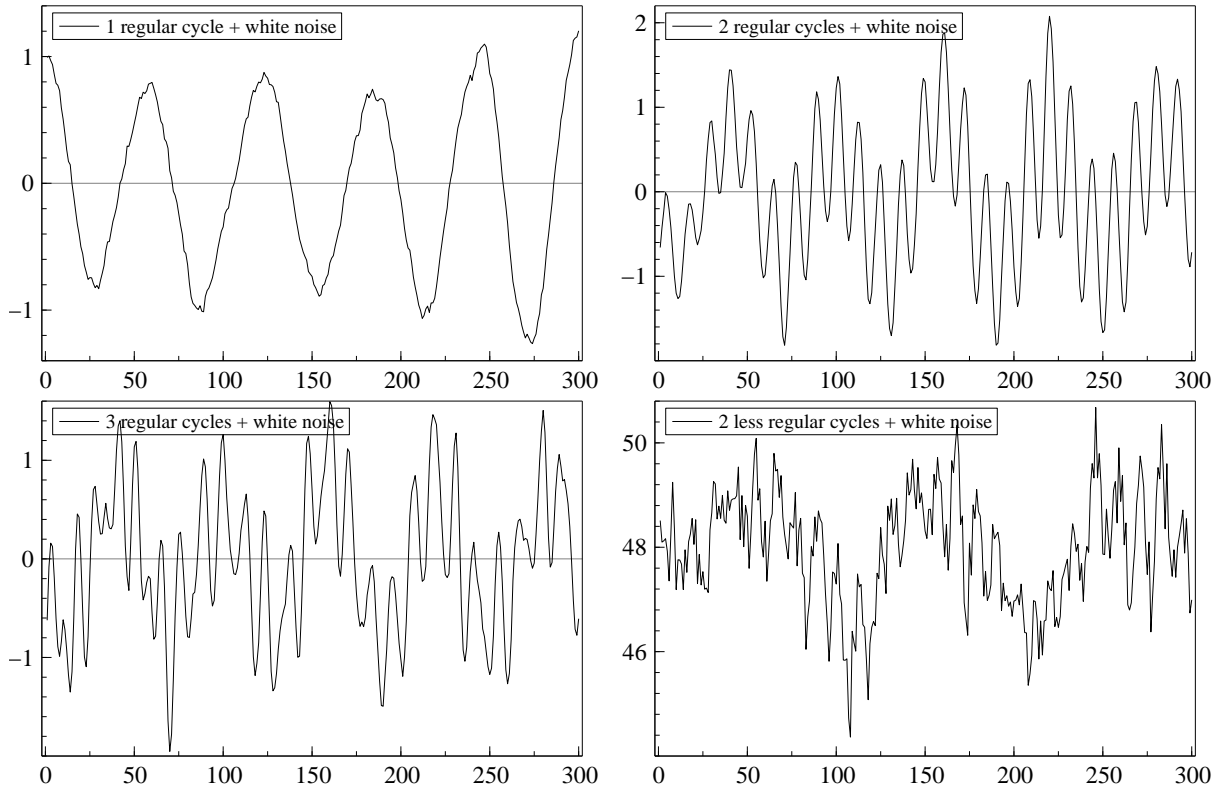


Figure 4.1: Artificial data with smooth cycles on the first 3 panels and less regular cycles on the bottom-right panel.

Table 4.1: Posterior model probabilities for artificial data example.

simulated model #cycles	#cycles in estimated model				
	1	2	3	4	5
1	0.531	0.252	0.112	0.077	0.028
2	0.000	0.454	0.320	0.179	0.046
3	0.000	0.000	0.516	0.344	0.140

*Note: Posterior probabilities under equal prior probabilities.*

For each of the 3 series, we estimate 5 models, in which the number of cycles ranges from 1 to 5, using the MCMC procedure. From the posterior sample we have calculated the marginal likelihoods and resulting model probabilities, based on equal prior model probabilities of  $1/5$ . These are shown in table 4.1. For each data set, the ‘true’ model gets the highest posterior probability. Models with ‘too many’ cycles get less and less posterior probability when the number of cycles is increased. Thus, the tendency for overfitting is limited. More importantly, models which contain too few cycles are firmly rejected, thereby excluding potentially strange outcomes, resulting from a model that is



not flexible enough to accommodate the features of the data, to enter the final posterior results after weighing by posterior model probability.

A more pronounced posterior statement on the minimal number of cycles required for an adequate description of the dynamics in the data could be obtained in a natural way by attaching prior probabilities to the number of cycles. A researcher might for instance believe that mechanisms that could generate cyclical behavior in economic time series plausibly account for 1, 2 or 3 distinguishable cycles, whereas e.g. 10 cycles would be mere noise fitting. A straightforward prior that can describe such information is a geometrically declining sequence of prior model probabilities, up to some pre-specified maximum number of cycles, for instance

$$Pr(C = c) \propto \left(\frac{1}{2}\right)^c, \quad \text{for } c = 1, \dots, C_{\max}$$

Such a prior would downweigh models with a higher number of cycles but it will be unable to fully compensate the low marginal likelihoods of the models with too few cycles. It effectively favours the smallest model that adequately describes the data. Note that the number of cycles should never exceed  $\lfloor T/2 \rfloor$ , because for that number of cycles the Fourier decomposition would guarantee a perfect fit.

### Artificial data with less regular cycles

Although such a prior on the number of cycles would put less probability mass on larger models, they will still receive some weight. The question remains how a model with more cycles than strictly necessary would affect posterior inference, especially when the model probabilities provide a much less clear-cut indication of the ‘correct’ number of cycles. For the purpose of answering this, we investigate in more detail the case where the generated data contains 2 cycles and the estimated model has 4 cycles, 2 of which are actually superfluous but which catch some (non-smooth) coincidental cycles that arise in small samples of noise. We simulated a series of 300 observations with 2 cycles of 120 and 12 periods, respectively, with smaller damping factors than before. Also, the irregular term now accounts for 40% of the total variation of the series. The result is a more irregular cyclical pattern, as shown in the bottom-right panel of figure 4.1.

In the posterior densities of the frequency parameters  $\omega_c$  in figure 4.2 we observe an interesting phenomenon. The cycle of the lowest frequencies that is present in the data is sometimes picked up by the first cycle of the estimated model and sometimes by the second one. The higher frequency cycle is picked up by either the second or third cycle in the estimated model. The fourth cycle in the model seems to pick up some spurious cycles in a wide range of frequencies, consistent with the flat spectrum of white noise. The third cycle seems to alternate between the higher frequency cycle and noise fitting. A few observations can be based on the resulting multi-modal posterior densities. First, had one used Maximum Likelihood to estimate the model, then only one of the modes would be the MLE. The uncertainty of the asymptotic curvature-based variance estimators of the MLE would not take into account the existence of the other modes. These problems become even more apparent for the cycles of higher frequency, where the likelihood has many spikes, such that cycles at the frequencies that coincidentally match the noise in the

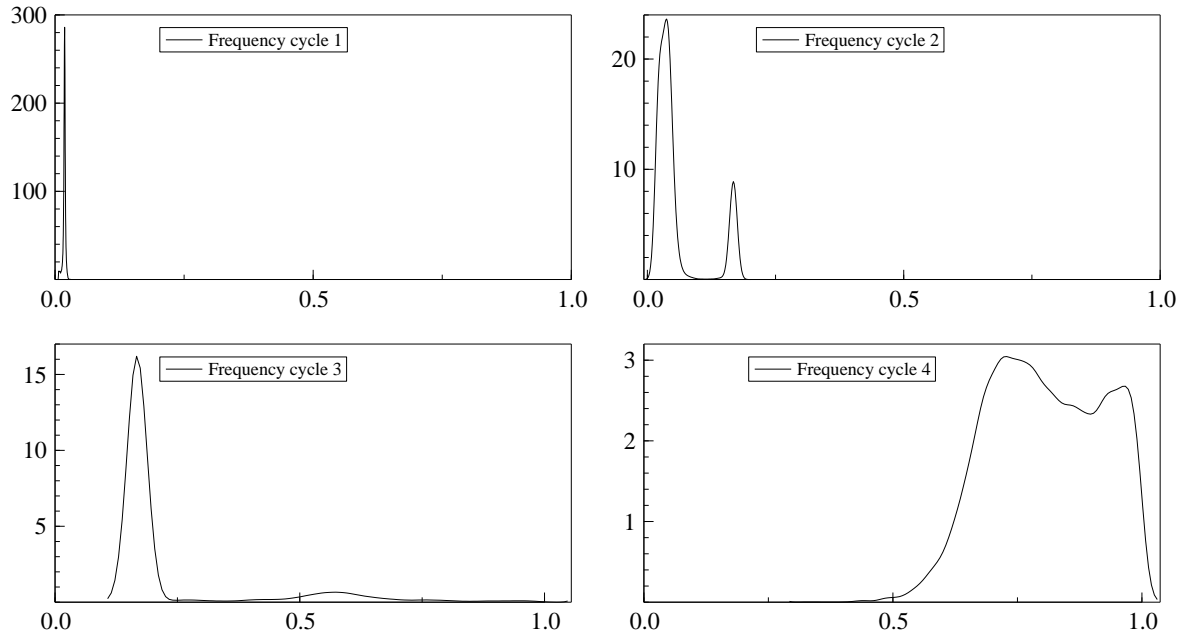


Figure 4.2: Bimodality in posterior density of frequency parameters. Simulated data contains 2 cycles, the estimated model has 4 cycles.

data are picked. Apart from that, any gradient-based optimization routine would become extremely sensitive to the starting values.

A second observation that can be made on the basis of the multi-modal densities concerns the problems that high-correlation MCMC chains will face. In our procedure, the sampler mixes slowly because the data augmentation inevitably introduces dependence between the 2 Gibbs steps. However, our candidate draw for the model parameters depends on the previous parameter draw only through the simulated states. And these states strongly depend on the observed data. If one uses for example a random walk Metropolis step or any other candidate that directly depends on the previous parameter draw, one essentially confines the sampling process to a localized search. In this way it becomes far more likely that other modes of the posterior density will be entirely missed.

So, although the posterior densities may look strange, they more accurately describe the inherent uncertainty on the number and frequency of the cycles than the MLE would or what a more correlated MCMC procedure would indicate. The researcher is seldom primarily interested in just the posterior distribution of the model parameters, but in more interesting functions of parameters and data. This model class involves cycles and hence it is natural to investigate some of its frequency domain properties. In particular we try to estimate the frequency spectrum and see how it is influenced by the number of cycles in the model. In the subsequent sections many more possible quantities of interest are reported.

Since we know how these data have been generated, we can derive the theoretical frequency spectrum from the parameters of the process. It is a standard result that the

spectrum for the cycle component (4.2c) at frequencies  $x \in [0, \pi]$  is given by

$$(2\pi)^{-1} \frac{1 + \rho^2 - 2\rho \cos \omega\pi \cos x}{1 + \rho^4 + 4\rho^2 \cos^2 \omega\pi - 4\rho(1 + \rho^2) \cos \omega\pi \cos x + 2\rho^2 \cos 2x} \sigma_\kappa^2,$$

see e.g. Harvey (1993, Ch.6), and by  $(2\pi)^{-1} \sigma_\xi^2$  for the irregular term. In the model the disturbances are assumed to be uncorrelated over the components. Therefore, all cross-spectra are zero and the spectrum of  $y_t$  can simply be computed as the sum of the spectra of its components. For this case of 2 cycles, this is shown in the top-left panel of figure 4.3. The theoretical spectrum clearly shows 2 peaks corresponding to the frequencies of the 2 cycles. The second panel displays the sample spectrum which is calculated from the sample autocorrelations of the simulated data. It is a well-known result that the accuracy of the estimated sample spectrum does not increase with sample size. This accounts for the non-smooth behavior of the sample spectrum. The top-right panel shows a smoothed version of the sample spectrum (based on a Parzen window).

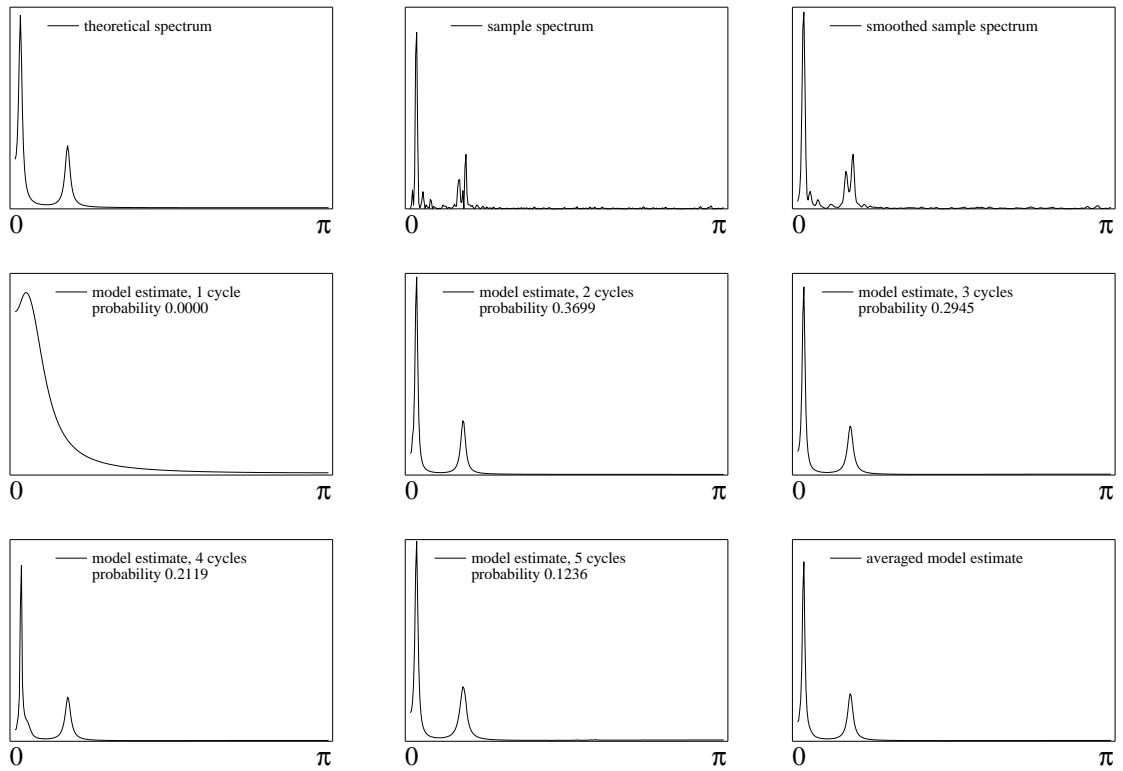


Figure 4.3: Theoretical, sample and estimated spectra for the artificial data.

The resulting graph is still not fully satisfactory as it gives the impression that there are 2 peaks around the frequency of the second cycle. Alternatively, one could calculate a parametric estimate of the spectrum. A common approach consists of fitting an ARMA model to the data and calculating its induced spectrum. The results of a parametric estimate based on our cycle decomposition procedure are reported on the middle and bottom rows of panels in figure 4.3. For each draw from the posterior density we have

calculated the induced power spectrum. At each frequency the posterior sample median of the spectrum is reported, but the full distribution of the spectrum at each frequency can be estimated from the sample output. In this way we can fully account for the parameter uncertainty.

The graph clearly shows that the model with only 1 cycle produces 1 very broad peak in an effort to encompass the 2 sharp peaks from the true process. This model is clearly misspecified, but its posterior probability is virtually zero. The model with 2 cycles accurately recovers the shape of the theoretical spectrum. The models with 3, 4 and 5 cycles yield very similar spectra. For a single draw from the posterior density, the superfluous cycles in these models give a third, fourth and fifth peak in the spectrum. The uncertainty of these cycles is so large that their peaks are not very sharp and they occur on different frequencies for different draws such that on average these superfluous peaks are spread out over the entire frequency range. In practice, basic visual inspection of the implied spectra could provide a hint of model adequacy. The bottom-right panel displays the estimated spectrum using a weighted average over the models using the calculated model probabilities. In this way we have not only fully accounted for the parameter uncertainty, but also for the model uncertainty within this class of cyclical models. Also, the parametric spectrum estimate recovers the theoretical spectrum more accurately than the nonparametric versions.

Summarizing, we find that a model with too few cycles is firmly rejected by our procedure. Most probability is generally attached to the ‘correct’ model. Model with ‘too many’ cycles are given some posterior probability, but this overfitting is quite harmless as the extra cycles are very uncertain and taking the uncertainty into account by averaging over the posteriors draws tends to cancel out their effect. Hence, their effect on posterior quantities tends to be negligible. Our procedure thus seems to be robust for misspecification.

### 4.3 US Industrial Production

Industrial production is a prominent cyclical sector of the economy. Therefore, Industrial Production seems a suitable series to investigate using our decomposition procedure. We shall investigate the durations of the growth cycles and whether the average growth rate has changed over time. We also produce a forecast incorporating all parameter and model uncertainty.

For modelling and forecasting monthly US industrial production Kawasaki and Franses (2004) prefer a structural time series model containing an  $I(2)$  random walk, a seasonal component and an irregular term. They argue that this specification comes close to the well-known airline model, which is often found useful to describe seasonal time series and combines a first-difference operator with a seasonal difference operator. Our focus will be on growth cycles at frequencies corresponding to business cycles and permanent effects and we therefore want to exclude most of the cyclical dynamics below the seasonal frequencies. As Kawasaki and Franses (2004) also find for this series that seasonality is non-stationary, we proceed by transforming the data to 12-month growth rates. The non-seasonal unit root that remains is captured by our  $I(1)$  trend component.

Table 4.2: Posterior model probabilities for US Industrial Production.

trend	#cycles in estimated model					total
	1	2	3	4	5	
no	0.000	0.000	0.005	0.052	0.035	0.092
yes	0.000	0.000	0.000	0.521	0.386	0.908

*Note: Posterior probabilities under equal prior probabilities.*

We have run the MCMC procedure for our model with monthly year-on-year growth rates of US Industrial production data from 1960:1 to 2003:10, obtained from Economagic. For the model with one cycle 300 iterations/second of the posterior simulator are achieved on current PC hardware.<sup>1</sup> We obtain an acceptance rate of 84%. In our experiments we have found that the computing times scale approximately linearly in the number of observations and quadratically in the number of cycles in the model.

The model probabilities are reported in table 4.2. The models with a stochastic trend component together get 91% posterior probability, providing evidence for non-stationary behavior of the Industrial Production growth rate. At least 4 cycles are required for an adequate description of the dynamics. In the favored model with a trend and 4 cycles the trend is assigned 17% of the overall variance and the cycles get respectively 53%, 26%, 17% and 4% of the remaining variance. The reported values have been calculated as the posterior means of  $\nu$  and  $\lambda_1$  to  $\lambda_4$ . In the posterior sampler these values are sampled simultaneously with the frequency parameters and therefore some caution is in order when analyzing variance shares and frequencies independent of each other. The implied posterior distribution of the spectral density provides a more integrated approach.

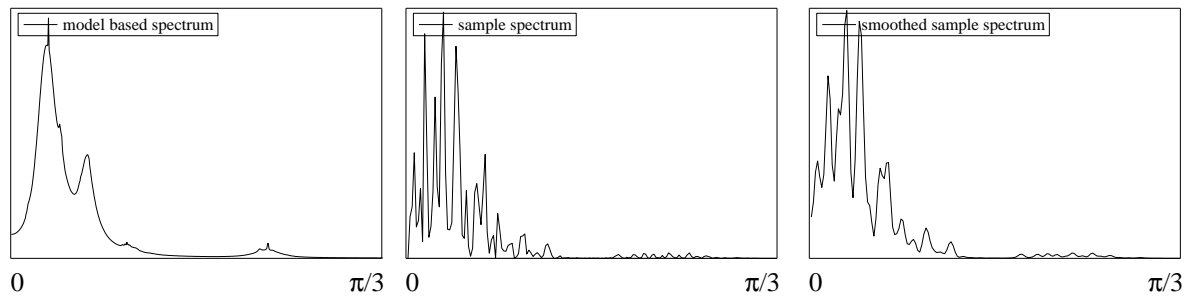


Figure 4.4: Sample spectra and model-based spectrum of stationary components for US industrial production data.

The posterior mean of the estimated spectrum of the stationary components is plotted in figure 4.4. It has 2 easily recognizable peaks at frequencies 0.102, and 0.217 corresponding to periods of oscillation of 62 and 29 months. There are also 2 smaller peaks at 0.329 and 0.731, describing cycles with cycle lengths of 19 and 9 months. Of

<sup>1</sup>All computations were carried out using Ox, see Doornik (1999) and the SsfPack package (Koopman et al. (1998)).

these cycles the first one is the most prominent one. It is assigned almost half of the total variance. Its 5 year period is in a range commonly regarded to be related to the business cycle.

Harding and Pagan (2002) prefer to specify the dating of business cycles in terms of the behaviour of its growth rates. In that perspective, we note that our method is informative on the behaviour of the growth rate at this frequency. The 5-year cycle provides a relatively clean signal whether growth is above or below average, while controlling for the cycles at higher frequencies. Alternatively, the 5-year cycle and the time-varying mean component together are sufficient to extract a signal of whether the cleaned growth rate is positive or negative.

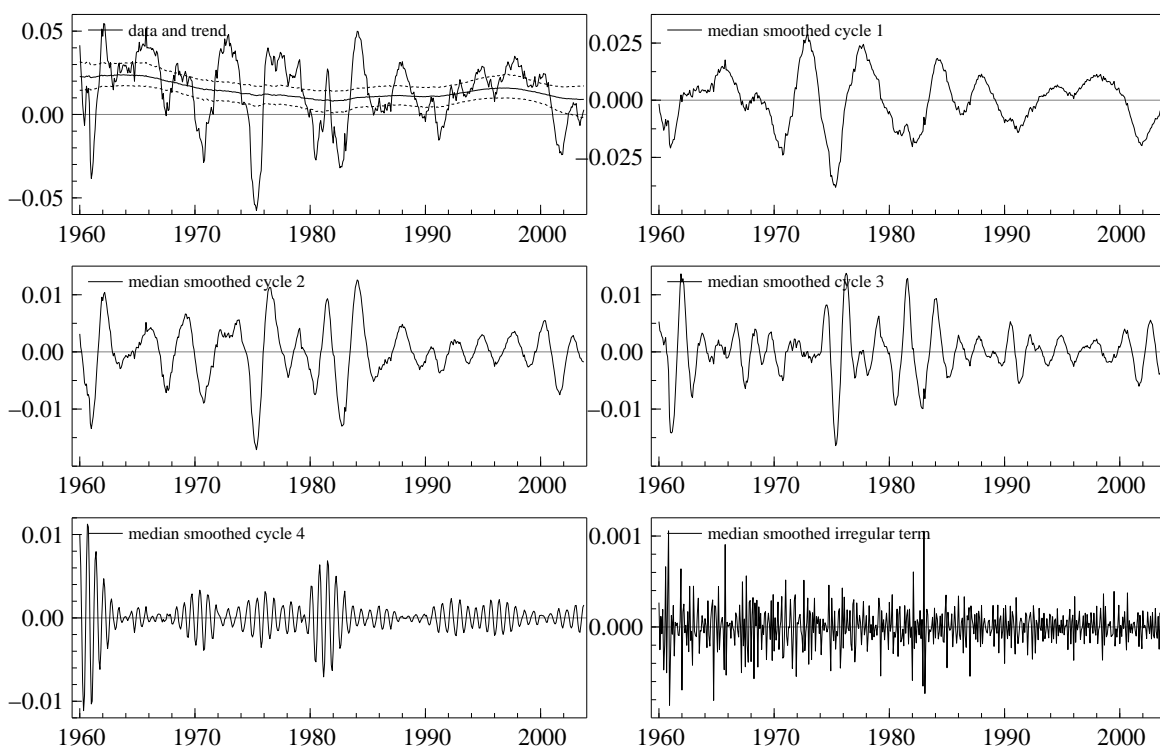


Figure 4.5: US Industrial Production, monthly data from 1960:1 to 2003:10. Posterior median and 5% and 95% quantiles of smoothed trend component. Also posterior medians of the other components.

The smoothed states from the preferred model with a stochastic trend and 4 cycles are shown in figure 4.5. The stochastic trend here has the interpretation of a time-varying mean growth rate. The time-varying mean component essentially remains after removing all cyclical components, which have mean 0 by construction, from the raw growth rates. We find that it has fallen from around 2.4% per year in the early 1960-s to around 1.0% per year in 1980. The 1990-s are characterized by a long period without recession during which the idea of the ‘new economy’ with permanent higher growth rates gained support. The higher growth rates in the 1990-s did not fit in the regular cyclical pattern, and therefore have been qualified as a permanent effect to the time-varying mean. Around

2000 this increase had already been neutralized again by subsequent changes in the mean  $\mu_t$ . The smoothed states of the first cycle show that in the 1990-s an expansion period was followed by a growth slowdown, but that it did not fully enter a contraction and it started a new expansion period in the second half of the 1990-s.

The cycle with highest frequency seems to capture some of the remaining seasonality in the data or some artifact of the transformation. It is interesting to note that the amplitude of this series seems to be higher in periods when the first cycle, which we associate with the business cycle, is negative. Although we do not formally test for it, it may be an indication that seasonality is stronger in periods of contraction than in expansions.

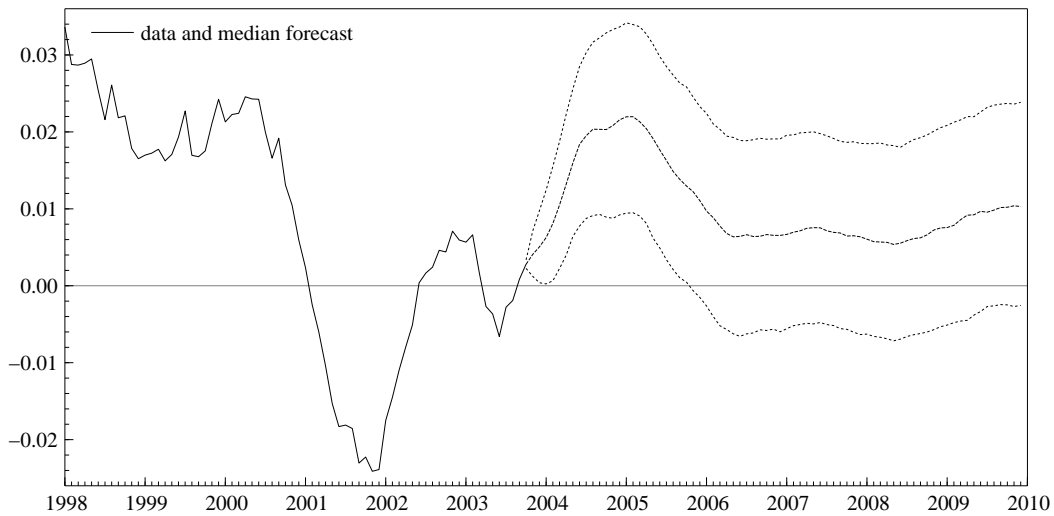


Figure 4.6: Posterior median and first and third quartiles of forecasts of US Industrial Production.

For each draw from the posterior density of the model parameters one can use standard state space methods for simulating an out-of-sample path of values for the states and the data, consistent with the in-sample data and the parameter values. By doing so for each parameter draw from the MCMC procedure and for each model using the model weights, one effectively obtains a full forecast density that incorporates all uncertainty arising from 5 sources, viz. the forecast uncertainty from the unknown future innovations, the estimation uncertainty of the model parameters, the uncertainty regarding the unobserved components, the model uncertainty regarding the number of cycles and the the model uncertainty regarding the inclusion of a trend component.

The median and first and third quartiles of such a forecast distribution, are given in figure 4.6.

## 4.4 US Unemployment

The unemployment rate is one of the most closely watched economic indicators. Its behavior is known to be asymmetric, typically exhibiting a long and slow decline followed

by a short but steep increase. We show how the model is able to deal with the asymmetry. For policy makers it is of interest to know how much of the unemployment is due to structural causes and how much can be attributed to cyclical and seasonal factors. We show how our model distinguishes between permanent and transitory components of unemployment. The variability in each is also of interest, especially whether the mean unemployment rate and the mean amplitude of the cycle are constant over time. We use monthly US unemployment rates starting 1959:01.

Table 4.3: Posterior model probabilities for the US Unemployment rate.

trend	#cycles in estimated model					total
	1	2	3	4	5	
no	0.000	0.006	0.064	0.067	0.264	0.401
yes	0.001	0.020	0.150	0.337	0.091	0.599

*Note: Posterior probabilities under equal prior probabilities.*

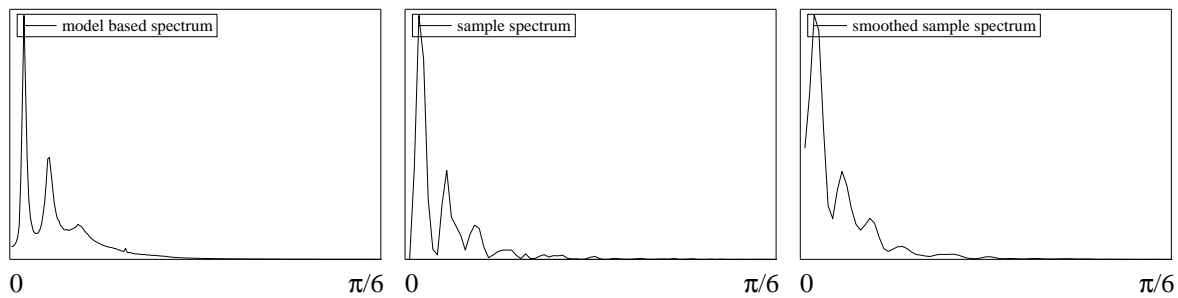


Figure 4.7: Sample spectra and model-based spectrum of stationary components for the US unemployment rate data.

The estimated model probabilities are reported in table 4.3. These indicate posterior odds of 60/40 in favor of a time-varying mean. The favored model contains a stochastic trend and 4 cycles. As before we calculated the implied posterior spectrum. The spectrum has 4 peaks as can be seen from figure 4.7. The first 3 peaks are more clearly discernable than the fourth. The frequencies of the modes are 0.018, 0.054, 0.094, and 0.162. These correspond to periods of 350, 117, 67, and 39 months, respectively. The posterior mean of the variance shares of the cycles are respectively 34%, 30%, 20% and 15%. The longest cycle of almost 30 years is below the usual business cycle frequencies. In the frequency domain it is closer to the zero frequency associated with the trend than it is to business cycle frequencies. As such it augments the trend for describing longer term phenomena in the data.

It is known that the superposition of a base cycle together with its higher order harmonics at exact integer multiples of the base frequency, can generate a large variety of interesting wave forms, such as block waves, triangular waves or sawtooth waves, even when the constituents are all pure sine waves. The sawtooth wave is an extreme example



of an asymmetric shape, where a single cycle consists of a gradual increase, followed by a instantaneous drop. The exact multiples in the periodicities keep the cycles synchronized, such that a certain phase of the base cycle always coincides with a certain recurring phase of one or more of the other cycles. In the context of stochastic cycles it emerges that such a mechanism is at action in the estimated model for the unemployment data, where the asymmetry is mimicked by the cycles. In particular the cycles peak at the same time.

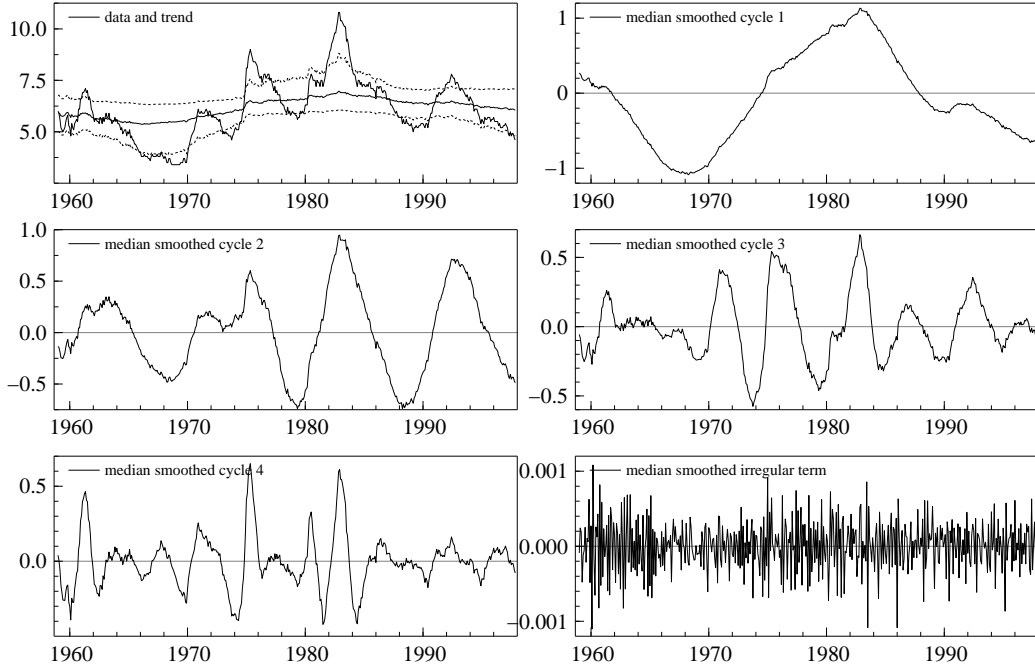


Figure 4.8: Smoothed states for the US unemployment rate data.

The smoothed states for the model with highest posterior probability, i.e. containing a trend and 4 cycles, are plotted in figure 4.8. The unobserved time-varying mean provided by the stochastic trend component measures the permanent part of the unemployment rate. It varies only slowly over time and has a value of 6.8% at the end of the observation period. When considered together with the longest cycle, the value becomes 6.3%.

The latent amplitude of the  $c$ -th stochastic cycle at time  $t$  is given by  $\sqrt{\psi_{ct}^2 + \psi_{ct}^{*2}}$ . The total latent amplitude, which is given by the sum of the amplitudes over the cycles, is the potential deviation when all cycles are simultaneously at their maximum value (or all at their minimum, but, of course, the unemployment rate cannot take on negative values). The amplitude thus serves as a measure of cyclical variability. Figure 4.9 shows the median and the 5% and 95% quantiles of the posterior distribution of the total smoothed latent amplitudes of the cyclical components at each point in time. The striking feature in the amplitudes are the peaks that occur only when the original data series switches from increasing to decreasing and not when the unemployment rate switches from decreasing to increasing. Presumably, the number of harmonics is sufficient to capture the latter transition but it is insufficient for describing the richer harmonics needed to describe the

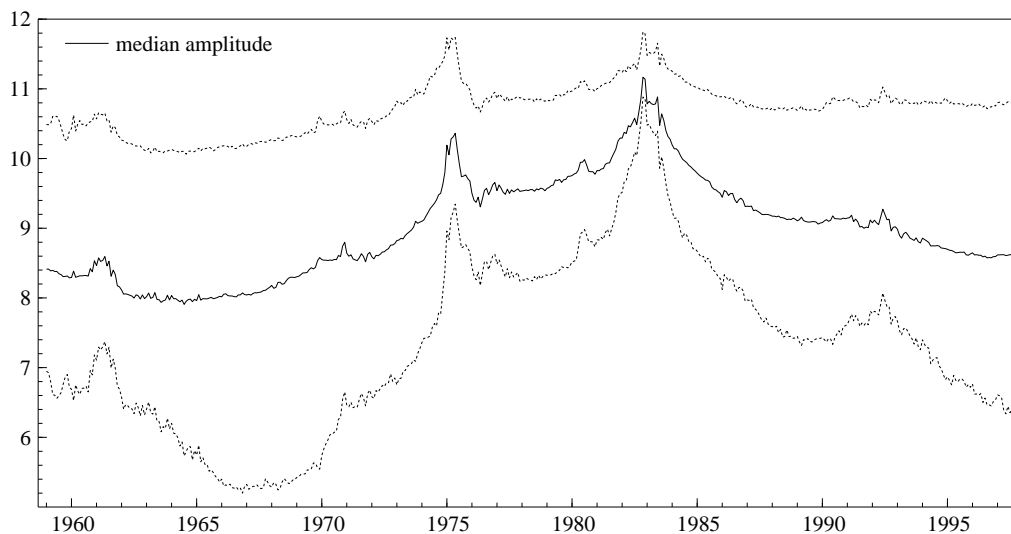


Figure 4.9: Smoothed joint amplitude of all cycles for the US unemployment rate data.

sudden downturn of the unemployment rates. In that case the flexibility of the stochastic cycles suffices to fit the harmonic dynamics.

We remark that there are alternative ways to capture asymmetry. An entire literature has emerged on (Markov) regime-switching models since Hamilton (1989), see e.g. Paap and Van Dijk (2003) for some recent Bayesian work. Also the quasi-periodic model of Kitagawa and Gersch (1996) can describe asymmetry. In their model a composition of harmonic trigonometric terms are modulated by treating their phase as a time-varying parameter. A third route is taken by Cogley, Morozov, and Sargent (2003) in which time-varying parameters in a VAR structure cause the spectral properties of the model to evolve over time. In our model an extension with time-variation of the parameter  $\rho_c$ , which controls the smoothness and mean-reversion of a cycle, could potentially also generate asymmetry using only a single cycle. This is topic for further research.

Another tendency that can be observed from figure 4.9 is the decline in amplitude that has occurred since 1985 to the end of the observation period back to the levels of the 1960-s. We view this as an indication that the unemployment rate has become less cyclical.

## 4.5 Summary and Conclusions

We have introduced a class of models, which consist of a stochastic trend, an *unknown* number of non-deterministic cyclical components and an irregular term. This specification leads to a very flexible functional form, which can capture a wide range of dynamics in the data. A MCMC-based Bayesian procedure is provided for sampling from the posterior distribution of the model parameters. Based on these samples, the posterior model probabilities can be calculated, indicating the appropriate number of cycles in the model. The parameter uncertainty and the uncertainty on the number of cycles can be

fully taken into account in the subsequent inference. Prior specification is facilitated by a reparametrization and an efficient candidate generator for the model parameters of unobserved cycle components accelerates the MCMC-scheme.

Simulated data were used to illustrate model choice, model averaging, and that noise fitting is avoided. We have provided estimates of the cycles in US Industrial Production, extracted the evolution of the average growth rate over time and calculated forecast densities that incorporate all uncertainty originating from several sources. The asymmetry of the cycles in the US Unemployment rate data are adequately described by cyclical components at some selected higher order harmonic frequencies. Also, the underlying evolving mean of the unemployment rate and the amplitude of the cyclical variability were calculated.

A lot of further research in this area remains to be explored, including the implications of the flexibility of this approach for e.g. real time extraction of the structural component in a trend-cycle decomposition, the dating of business cycles and other empirical applications. Possible model extensions to consider are multivariate analysis and explicit modelling of the synchronization mechanism between cycles.

## APPENDICES

### 4.A Posterior sampler for a single cycle component

Consider  $T$  observations from one of the cycle components, i.e.:

$$\begin{aligned} \begin{pmatrix} \psi_t \\ \psi_t^* \end{pmatrix} &= \rho \begin{pmatrix} \cos \omega \pi & \sin \omega \pi \\ -\sin \omega \pi & \cos \omega \pi \end{pmatrix} \begin{pmatrix} \psi_{t-1} \\ \psi_{t-1}^* \end{pmatrix} + \begin{pmatrix} \kappa_t \\ \kappa_t^* \end{pmatrix} \\ &= \begin{pmatrix} a & b \\ -b & a \end{pmatrix} \begin{pmatrix} \psi_{t-1} \\ \psi_{t-1}^* \end{pmatrix} + \begin{pmatrix} \kappa_t \\ \kappa_t^* \end{pmatrix}, \quad \text{for } t = 1, \dots, T \end{aligned}$$

with  $a = \rho \cos \omega \pi$  and  $b = \rho \sin \omega \pi$  and inversely  $\rho = \sqrt{a^2 + b^2}$  and  $\omega = \frac{1}{\pi} \arctan \frac{b}{a}$ . We define  $\Psi = (\psi_1 \dots \psi_T)'$ ,  $\Psi_{-1} = (\psi_0 \dots \psi_{T-1})'$ , and  $\kappa = (\kappa_1 \dots \kappa_T)'$ , and also  $\Psi^*$ ,  $\Psi_{-1}^*$  and  $\kappa^*$  in an analogous way. In matrix notation the model can be expressed as

$$\begin{pmatrix} \Psi & \Psi^* \end{pmatrix} = \begin{pmatrix} \Psi_{-1} & \Psi_{-1}^* \end{pmatrix} \begin{pmatrix} a & -b \\ b & a \end{pmatrix} + \begin{pmatrix} \kappa & \kappa^* \end{pmatrix}$$

and by taking the vec at both sides we get

$$\begin{aligned} \begin{pmatrix} \Psi \\ \Psi^* \end{pmatrix} &= (I_2 \otimes (\Psi_{-1} \quad \Psi_{-1}^*)) \text{vec} \begin{pmatrix} a & -b \\ b & a \end{pmatrix} + \begin{pmatrix} \kappa \\ \kappa^* \end{pmatrix} \\ &= (I_2 \otimes (\Psi_{-1} \quad \Psi_{-1}^*)) \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 0 & -1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} + \begin{pmatrix} \kappa \\ \kappa^* \end{pmatrix} \\ &= \begin{pmatrix} \Psi_{-1} & \Psi_{-1}^* \\ \Psi_{-1}^* & -\Psi_{-1} \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} + \begin{pmatrix} \kappa \\ \kappa^* \end{pmatrix} \end{aligned}$$

Taking  $y = \begin{pmatrix} \Psi \\ \Psi^* \end{pmatrix}$  and  $X = \begin{pmatrix} \Psi_{-1} & \Psi_{-1}^* \\ \Psi_{-1}^* & -\Psi_{-1} \end{pmatrix}$ , this takes algebraically the form of a simple linear regression

$$y = X\beta + \varepsilon, \quad \varepsilon \sim N(0, \sigma_\kappa^2 I_{2T})$$

with the OLS estimator given by  $\hat{\beta} = (X'X)^{-1}X'y$  given by

$$\frac{1}{\Psi_{-1}'\Psi_{-1} + \Psi_{-1}^*\Psi_{-1}^*} \begin{pmatrix} \Psi_{-1}'\Psi + \Psi_{-1}^*\Psi^* \\ \Psi_{-1}^*\Psi - \Psi_{-1}\Psi^* \end{pmatrix}$$

The Jacobian of the transformation from  $(a, b, \sigma_\kappa^2)$  to  $(\rho \cos \omega \pi, \rho \sin \omega \pi, (1 - \rho^2)\sigma_\psi^2)$  is

$$\left| \frac{\partial (a \quad b \quad \sigma_\kappa^2)'}{\partial (\rho \quad \omega \quad \sigma_\psi^2)} \right| = \begin{vmatrix} \cos \omega \pi & -\rho \pi \sin \omega \pi & 0 \\ \sin \omega \pi & \rho \pi \cos \omega \pi & 0 \\ -2\rho \sigma_\psi^2 & 0 & 1 - \rho^2 \end{vmatrix} = (1 - \rho^2)\rho\pi$$

## 4.B Technical Details of the Posterior Simulator

The Metropolis-Hastings (M-H) step within the Gibbs sampler in subsection 4.2.4 produces a draw from the posterior density  $p(\theta|\alpha, y)$ . This density is proportional to the product of our prior  $p(\theta)$  from subsection 4.2.3 and the likelihood  $L(\alpha, y|\theta)$  for model (4.2) in which the states can now be treated as observables. The likelihood (and hence also the posterior) factorizes into components as

$$\begin{aligned} L(\alpha, y|\theta) &= L(y, \alpha_2, \dots, \alpha_T|\alpha_1, \theta) L(\alpha_1|\theta) \\ &= L(y|\alpha, \sigma_\xi^2) L(\mu_2, \dots, \mu_T|\mu_1, \sigma_\mu^2) \\ &\quad \times \prod_{c=1}^C L(\psi_{c2}, \dots, \psi_{cT}, \psi_{c2}^*, \dots, \psi_{cT}^*|\psi_{c1}, \psi_{c1}^*, \sigma_c^2, \omega_c, \rho_c) L(\psi_{c1}, \psi_{c1}^*|\sigma_c^2, \omega_c, \rho_c), \end{aligned}$$

which successively are proportional to inverted-gamma (IG) densities on  $\sigma_\xi^2$  and  $\sigma_\eta^2 \equiv \frac{6}{T}\sigma_\mu^2$ , to a normal-inverted gamma (NIG) density on the parameters of each of the cycles  $c$  (using the transformation from the first appendix) and to a bivariate normal density for the initial conditions of each cycle, with  $L(\psi_{c1}, \psi_{c1}^*|\sigma_c^2, \omega_c, \rho_c) \propto \sigma_c^{-2} \exp(-\frac{1}{2}(\psi_{c1}^2 + \psi_{c1}^{*2})/\sigma_c^2)$ . With the exception of the initial conditions, which do not combine in a simple way with the reparameterized cycle parameters, each of these distributions is very straightforward to deal with. However, the structure that the prior  $p(\theta)$  provides to our analysis comes at a small cost. It neither is conjugate to the normal and IG distributions nor can it be factorized conveniently. In the M-H step we therefore use a more practical pseudo prior  $q(\theta)$  instead of  $p(\theta)$  and also exclude the initial conditions in the candidate density  $q(\theta|\alpha, y) \propto L(y, \alpha_2, \dots, \alpha_T|\alpha_1, \theta)q(\theta)$ . The acceptance probability deals with the discrepancy between  $p(\theta)L(\alpha_1|\theta)$  and  $q(\theta)$ .

The simplest choice for  $q(\theta)$  would consist of independent non-informative priors, i.e.  $q(\theta) = q(\sigma_\eta^2)q(\sigma_\xi^2)\prod_{c=1}^C q(\sigma_c^2, \omega_c, \rho_c)$ , with  $q(\sigma_\eta^2) \propto \sigma_\eta^{-2}$  and  $q(\sigma_\xi^2) \propto \sigma_\xi^{-2}$ . For the parameters of the  $c$ -th cycle the prior is chosen to be consistent with a non-informative prior  $q(\beta_c, \sigma_{\kappa_c}^2) \propto \sigma_{\kappa_c}^{-2}$  on the transformed parameters  $\beta_c$  and  $\sigma_{\kappa_c}^2$  from the first appendix, such that

$$q(\sigma_c^2, \omega_c, \rho_c) = q(\beta_c, \sigma_{\kappa_c}^2) \left| \frac{\partial (\beta_c \quad \sigma_{\kappa_c}^2)'}{\partial (\rho_c \quad \omega_c \quad \sigma_c^2)} \right| = \frac{1}{(1 - \rho_c^2)\sigma_c^2} (1 - \rho_c^2) \rho_c \pi = \frac{\rho_c \pi}{\sigma_c^2}.$$

One might try to improve on this choice of  $q(\theta)$  by using a (conjugate) prior that captures more of the structure of  $p(\theta)$ , for example by matching some of its moments. In our experience the sampler runs fine without this extra effort. Using these candidate priors and defining  $X_c$ ,  $\beta_c$  and  $\varepsilon_c$  similar to  $X$ ,  $\beta$  and  $\varepsilon$  from the previous section, the M-H step is implemented as

1. (a) generate candidate  $\sigma_\eta^2$  from  $q(\sigma_\eta^2|\eta)$  as  $\frac{\sum_{t=2}^T \eta_t^2}{\sigma_\eta^2}|\eta \sim \chi_{T-1}^2$ , with  $\eta_t = \Delta\mu_t$
- (b) generate candidate  $\sigma_\xi^2$  from  $q(\sigma_\xi^2|\xi)$  as  $\frac{\sum_{t=1}^T \xi_t^2}{\sigma_\xi^2}|\xi \sim \chi_T^2$ , with  $\xi_t = y_t - \sum \psi_{ct} - \mu_t$
- (c) generate  $\sigma_{\kappa_c}^2$  from  $q(\sigma_{\kappa_c}^2|\Psi_c, \Psi_c^*)$  as  $\frac{\sum_{t=1}^T \varepsilon_{ct}^2}{\sigma_{\kappa_c}^2}|\Psi_c, \Psi_c^* \sim \chi_{T-2}^2$  (for  $c = 1, \dots, C$ )

- (d) generate  $\beta_c$  from  $q(\beta_c|\sigma_{\kappa_c}^2, \Psi_c, \Psi_c^*)$  as  $\beta_c|\sigma_{\kappa_c}^2, \Psi_c, \Psi_c^* \sim N(\hat{\beta}_c, \sigma_{\kappa_c}^2 (X_c' X_c)^{-1})$  (for  $c = 1, \dots, C$ )
- (e) map  $(\beta_c, \sigma_{\kappa_c}^2)$  into candidate  $(\sigma_c^2, \omega_c, \rho_c)$  (for  $c = 1, \dots, C$ )
- 2. with probability  $AP(\theta^*|\theta^{(-1)}, \alpha, y)$  accept the candidate  $\theta^* = (\sigma_\eta^2, \sigma_\xi^2, \sigma_1^2, \dots, \sigma_C^2, \omega_1, \dots, \omega_C, \rho_1, \dots, \rho_C)$  or keep the previous parameter value  $\theta^{(-1)}$  otherwise.

The acceptance probability is given by

$$AP(\theta^*|\theta^{(-1)}, \alpha, y) = \min \left\{ 1, \frac{p(\theta^*|\alpha, y)}{q(\theta^*|\alpha, y)} \bigg/ \frac{p(\theta^{(i-1)}|\alpha, y)}{q(\theta^{(i-1)}|\alpha, y)} \right\}$$

and this can be simplified by noting that  $\frac{p(\theta|\alpha, y)}{q(\theta|\alpha, y)} = \frac{L(\alpha_1|\theta)p(\theta)}{q(\theta)}$ . We already specified  $L(\alpha_1|\theta)$  and  $q(\theta)$ . In (4.3) we have specified the prior distributions in terms of the parameterization  $\tilde{\theta}$ . Therefore,  $p(\theta)$  is given by  $p(\theta) = p(\tilde{\theta}) \left| \frac{\partial \tilde{\theta}}{\partial \theta} \right|$ . The Jacobian determinant can be expressed as  $\left| \frac{\partial \theta}{\partial \tilde{\theta}} \right|^{-1} = (1 - \nu)^{-C} (\sigma^2)^{-(C+1)}$  because with  $\sigma^2 = \sigma_\xi^2 + \sigma_\mu^2 + \sum \sigma_c^2$ ,  $\nu = \frac{\sigma_\mu^2}{\sigma^2}$ ,  $\lambda_c = \frac{\sigma_c^2}{(1-\nu)\sigma^2}$  we can derive that

$$\begin{aligned} & \left| \frac{\partial (\sigma_1^2 \dots \sigma_C^2 \sigma_\xi^2 \sigma_\mu^2)'}{\partial (\lambda_1 \dots \lambda_C \sigma^2 \nu)} \right| \\ &= \begin{vmatrix} (1-\nu)\sigma^2 & & 0 & (1-\nu)\lambda_1 & -\lambda_1\sigma^2 \\ & \ddots & & \vdots & \vdots \\ 0 & & (1-\nu)\sigma^2 & (1-\nu)\lambda_C & -\lambda_C\sigma^2 \\ -(1-\nu)\sigma^2 & \dots & -(1-\nu)\sigma^2 & (1-\nu)(1-\sum \lambda_c) & -(1-\sum \lambda_c)\sigma^2 \\ 0 & \dots & 0 & \nu & \sigma^2 \end{vmatrix} \\ &= \begin{vmatrix} P & R \\ S & Q \end{vmatrix} = |P||Q - SP^{-1}R| = |P|\sigma^2 = (1-\nu)^C (\sigma^2)^{C+1}. \end{aligned}$$

Here we have used that the inverse transformation from  $(\lambda_1 \dots \lambda_C \sigma^2 \nu)$  to  $(\sigma_1^2 \dots \sigma_C^2 \sigma_\xi^2 \sigma_\mu^2)$  can be expressed as  $\sigma_c^2 = (1-\nu)\lambda_c\sigma^2$ ,  $\sigma_\xi^2 = (1-\nu)(1-\sum \lambda_c)\sigma^2$ , and  $\sigma_\mu^2 = \nu\sigma^2$ .

Note that the ordering of the frequencies is not specified in the candidate. In our case the strong dependence of the conditional candidate  $q(\theta|\alpha, y)$  on the augmented unobserved states to a great extent preclude the generation of an infeasible candidate. If they do occur they are automatically rejected as their acceptance probability will always equal zero. In related work on Bayesian analysis of models with latent processes under ordering constraints, though only taking values in a discrete state space, Frühwirth-Schnatter (2001) employs a permutation sampler which adds a reordering pass to ensure a unique labeling and to improve the acceptance rate of the MCMC sampler.

## Chapter 5

# Forecast Accuracy and Economic Gains from Bayesian Model Averaging using Time Varying Weights

*Chapter 5 is based on Hoogerheide, Kleijn, Ravazzolo, van Dijk, and Verbeek (2010).*

## 5.1 Introduction

When an extensive set of forecasts of some future economic event is available, decision makers usually attempt to discover which is the best forecast, then accept this and discard the other forecasts. However, the discarded forecasts may have some independent valuable information and including them in the forecasting process may provide more accurate results. An important explanation is related to the fundamental assumption that in most cases one cannot identify *a priori* the exact true economic process or the forecasting model that generates smaller forecast errors than its competitors. Different models may play a – possibly temporary – complementary role in approximating the data generating process. In these situations, forecast combinations are viewed as a simple and effective way to obtain improvements in forecast accuracy.

Since the seminal article of Bates and Granger (1969) several papers have shown that combinations of forecasts can outperform individual forecasts in terms of loss functions. For example, Stock and Watson (2004) find that for predicting output growth in seven countries forecast combinations generally perform better than forecasts based on single models. Marcellino (2004) has extended this analysis to a large European data set with broadly the same conclusion. However, several alternative combination schemes are available and it is not clear which is the best scheme, either in a frequentist or Bayesian framework. For example, Hendry and Clements (2004) and Timmermann (2006) show that simple combinations<sup>1</sup> often give better performance than more sophisticated approaches. Further, using a frequentist approach, Granger and Ramanathan (1984) propose the use of coefficient regression methods, Hansen (2007) introduces a Mallows' criterion, which can be minimized to select the empirical model weights, and Terui and Van Dijk (2002) generalize the least squares model weights by reformulating the linear regression model as a state space specification where the weights are assumed to follow a random walk process. Guidolin and Timmermann (2009) propose a different time varying weight combination scheme where weights have regime switching dynamics. Stock and Watson (2004) and Timmermann (2006) use the inverse mean square prediction error (MSPE) over a set of the most recent observations to compute model weights. In a Bayesian framework, Madigan and Raftery (1994) revitalize the concept of Bayesian model averaging (BMA) and apply it in an empirical application dealing with Occam's Window. Recent applications suggest its relevance for macroeconomics (Fernández et al., 2001 and Sala-I-Martin et al., 2004). Strachan and Van Dijk (2008) compute impulse response paths and effects of policy measures using BMA in the context of a large set of vector autoregressive models. Geweke and Whiteman (2006) apply BMA using predictive likelihoods instead of marginal likelihoods.

This paper contributes to the research on forecast combinations by investigating several Bayesian combination schemes. We propose three schemes that allow for parameter uncertainty, model uncertainty and time varying model weights simultaneously. These

---

<sup>1</sup>Simple combinations are defined as combinations with model weights that do not involve unknown parameters to be estimated; arithmetic averages constitute a simple example. Complex combinations are defined as combinations that rely on estimating weights that depend on the full variance-covariance matrix and, possibly, allow for time varying model weights.



approaches can be considered Bayesian extensions of the combination scheme of Terui and Van Dijk (2002).

We provide two empirical illustrations. The results indicate that time varying model weight schemes outperform other averaging schemes in terms of predictive and economic gains. The first empirical example deals with forecasting the returns on the S&P 500 index by combining individual forecasts from four competing models. The first model assumes that a set of financial and macroeconomic variables that are related to the business cycle have explanatory power. The second model is based on the popular market saying “Sell in May and go away”, also known as the “Halloween indicator”, see for example Bouman and Jacobsen (2002). Low predictability of stock market return data is well documented, see for example Marquering and Verbeek (2004) and so is structural instability in this context, see for example Pesaran and Timmermann (2002) and Ravazzolo et al. (2008). The third and fourth model are (robust) stochastic volatility models. As an investor is particularly interested in the economic value of a forecasting scheme, we test our findings in an active short-term investment exercise, with an investment horizon of one month. The forecast combination schemes with time-varying model weights provide the highest economic gains. The second empirical example refers to forecasting US economic growth over the business cycle, where we consider combinations of forecasts from six well-known time series models: an autoregressive model, two random walk models (with and without drift), an error correction model and two (robust) stochastic volatility models. It suggests that time varying weighting schemes may provide an early indicator for recessions.

The contents of this paper are organized as follows. In Section 5.2 we describe the different forecast combination schemes. In Section 5.3 we give results from an empirical application to US stock index returns which show that forecast combinations give economic gains. In Section 5.4 we report results from macroeconomic forecasts using US GDP growth. Section 5.5 concludes.

## 5.2 Forecast combination schemes

Bayesian approaches have been widely used to construct forecast combinations, see for example Leamer (1978), Hodges (1987), Draper (1995), Min and Zellner (1993), and Strachan and Van Dijk (2008). In the Bayesian model averaging approach one derives the posterior density for any individual model and combines these to compute a predictive density of the event of interest. The predictive density accounts then for model uncertainty by averaging over the posterior probabilities of individual models. Since the output is a complete density, not only point forecasts but also distribution and quantile forecasts can be easily derived. We discuss four Bayesian forecast combination schemes. The first scheme is a standard approach known as Bayesian model averaging, the other three schemes obtain model weights as parameters to be estimated in linear and nonlinear regressions.

### 5.2.1 Scheme 1: Bayesian Model Averaging (BMA)

The predictive density of the variable  $y$  at time  $T + 1$ ,  $y_{T+1}$ , given the data up to time  $T$ ,  $D_T$ , is computed by averaging over the conditional predictive densities given the individual models with the posterior probabilities of these models as weights:

$$p(y_{T+1}|D_T) = \sum_{i=1}^n p(y_{T+1}|D_T, m_i)P(m_i|D_T) \quad (5.1)$$

where  $n$  is the number of individual models;  $p(y_{T+1}|D_T, m_i)$  is the conditional predictive density given  $D_T$  and model  $m_i$ ;  $P(m_i|D_T)$  is the posterior probability for model  $m_i$ . The conditional predictive density given  $D_T$  and model  $m_i$  is defined as:

$$p(y_{T+1}|D_T, m_i) = \int p(y_{T+1}|D_T, m_i, \theta_i)p(\theta_i|D_T, m_i)d\theta_i \quad (5.2)$$

where  $p(y_{T+1}|D_T, m_i, \theta_i)$  is the conditional predictive density of  $y_{T+1}$  given  $D_T$ , the model  $m_i$  and parameters  $\theta_i$ ;  $p(\theta_i|D_T, m_i)$  is the posterior density for parameters  $\theta_i$  in model  $m_i$ .

The posterior probability for model  $m_i$ ,  $P(m_i|D_T)$ , can be computed in several ways. Madigan and Raftery (1994) define it as:

$$P(m_i|D_T) = \frac{p(y_{1:T}|m_i)P(m_i)}{\sum_{j=1}^n p(y_{1:T}|m_j)P(m_j)} \quad (5.3)$$

where  $y_{1:T} = \{y_t\}_{t=1}^T$ ;  $P(m_i)$  is the prior probability for model  $m_i$ ; and  $p(y_{1:T}|m_i)$  is the marginal likelihood for model  $m_i$  given by:

$$p(y_{1:T}|m_i) = \int p(y_{1:T}|\theta_i, m_i)p(\theta_i|m_i)d\theta_i \quad (5.4)$$

with  $p(\theta_i|m_i)$  the prior density for the parameters  $\theta_i$  in model  $m_i$ . The integral in equation (5.4) can be evaluated analytically in the case of linear models, but not for more complex forms. Chib (1995), for example, has derived a method to compute the expression also for nonlinear examples. Laplace methods can also be used, see for example Planas et al. (2008). A comparative study of Monte Carlo methods for marginal likelihood evaluation, among which importance sampling and bridge sampling, is given by Ardia et al. (2009).

Geweke and Whiteman (2006) propose a BMA scheme based on the idea that a model is as good as its predictions. The predictive density of  $y_{T+1}$  conditional on  $D_T$  has the same form as equation (5.1), but the posterior probability of model  $m_i$  conditional on  $D_T$  is now computed as:

$$P(m_i|D_T) = \frac{p(y_T|D_{T-1}, m_i)P(m_i)}{\sum_{j=1}^n p(y_T|D_{T-1}, m_j)P(m_j)} \quad (5.5)$$

where  $p(y_T|D_{T-1}, m_i)$  is the predictive likelihood for model  $m_i$ , e.g. the density derived by substituting the realized value  $y_T$  into the predictive density of  $y_T$  conditional on  $D_{T-1}$  given model  $m_i$ . Mitchell and Hall (2005) discuss the relation of the predictive likelihood to the Kullback-Leibler Information Criterion, and consequently to the

frequentist combination scheme based on recursive log-score weights, see for example Kascha and Ravazzolo (2008).

We apply BMA using (5.5) with  $p(y_T|D_{T-1}, m_i)$  replaced by its product over  $T - k$  observations  $p(y_{k+1}|D_k, m_i) \times \dots \times p(y_T|D_{T-1}, m_i)$ , where for increasing  $T$  we hold constant the length  $k$  of the ‘initial period’ of data  $D_k$  that are only used for deriving posterior distributions.<sup>2</sup> That is, for forecasts of  $y_{T+1}$  in later periods the predictive likelihoods and model weights are based on an expanding window of data. The densities  $p(y_t|D_{t-1}, m_i)$  are evaluated as follows. First, parameters  $\theta_i$  are simulated from the conditional distribution on  $D_{t-1}$ . Second, draws  $y_t$  are simulated conditionally on the  $\theta_i$  draws and  $D_{t-1}$ . Third, a kernel smoothing technique is used to estimate the density of  $y_t$  in model  $m_i$  at its realized value. The performance of alternative approaches for computing predictive likelihoods in our time varying model combination schemes is left as a topic for future research.

In all models, we specify uninformative proper priors for the parameters  $\theta_i$ . The use of predictive likelihoods rather than marginal likelihoods helps us to avoid the inference problems due to the Bartlett paradox.

### 5.2.2 Combination schemes using estimated regression coefficients as model weights

The next three combination schemes estimate the weights  $w_i$  of the models  $m_i$  ( $i = 1, \dots, n$ ) in regression form. We assume that the data  $y_t$  satisfy the linear equation

$$y_t = w_0 + \sum_{i=1}^n w_i y_{t,i} + u_t \quad u_t \sim N(0, \sigma^2) \text{ i.i.d.} \quad t = 1, 2, \dots, T \quad (5.6)$$

where  $y_{t,i}$  has the predictive density  $p(y_t|D_{t-1}, m_i)$  of  $y_t$  given  $D_{t-1}$  in model  $m_i$ . Clear differences with the BMA approach are that a constant term  $w_0$  is added, and that there is no restriction that all weights must be non-negative and adding to 1.<sup>3</sup> Therefore, the weights  $w_i$  ( $i = 1, \dots, n$ ) cannot be interpreted as model probabilities. Define the model weight vector  $w = (w_0, w_1, \dots, w_n)'$ . We propose three novel sampling algorithms for simulating model weight vectors  $w$  given the data  $y_{1:T}$  and the predictive densities  $p(y_t|D_{t-1}, m_i)$  ( $t = 1, \dots, T$ ).

#### Scheme 2: Model weights from Ordinary Least Squares in a linear model (LIN)

A set of model weight vectors  $w^s$  ( $s = 1, \dots, S$ ) is generated by simulating independently  $S$  sets of  $T \times n$  draws  $y_{t,i}^s$  from the predictive densities  $p(y_t|D_{t-1}, m_i)$  ( $t = 1, \dots, T$ ;  $i = 1, \dots, n$ ), and performing an Ordinary Least Squares (OLS) regression in the model

$$y_t = w_0 + \sum_{i=1}^n w_i y_{t,i}^s + u_t^s \quad u_t^s \sim N(0, \sigma^2) \quad t = 1, 2, \dots, T \quad (5.7)$$

<sup>2</sup>We choose  $k = 12$  for our applications involving monthly data.

<sup>3</sup>Granger and Ramanathan (1984) explain that the constant term must be added to avoid biased forecasts. They also conclude that this strategy is often more accurate than using restricted least squares weights.

for each simulated set  $s = 1, \dots, S$ . It is well-known that in a linear model as (5.7) the OLS estimator  $w^s$  is the posterior mean of  $w$  under a flat prior. The generated model weights  $w^s$  are used to combine draws  $y_{T+1,i}^s$  ( $i = 1, \dots, n$ ) from the predictive densities  $p(y_{T+1}|D_T, m_i)$  into ‘combined draws’  $\tilde{y}_{T+1}^s$ :

$$\tilde{y}_{T+1}^s = w_0^s + \sum_{i=1}^n w_i^s y_{T+1,i}^s \quad (5.8)$$

The median of  $\tilde{y}_{T+1}^s$  ( $s = 1, \dots, S$ ) is our point forecast  $\hat{y}_{T+1}$  for  $y_{T+1}$ , where the median is preferred over the mean because it is more robust to extreme draws. This approach can be considered as an extension of the idea of Granger and Ramanathan (1984) to combine point forecasts using weights that minimize a square loss function, to making use of Bayesian density forecasts. The model weights minimize the *distance* between the vector of observed values  $y_{1:T}$  and the space spanned by the constant vector and the vectors of ‘predicted’ values  $y_{1:T,i}^s$  ( $i = 1, \dots, n$ ).

The ‘combined draws’  $\tilde{y}_{T+1}^s$  are interpreted as draws from a ‘shrunk’ predictive density that aims at describing the central part of the predictive density, taking into account the parameter and model uncertainty.

The assumption that the error term  $u_t^s$  in (5.7) has constant variance  $\sigma^2$  and no serial correlation over  $t$ , and has a normal distribution, is arguably violated. However, violations of this assumption have no dire consequences for the performance of the proposed point forecast  $\hat{y}_{T+1}$ . Roughly stated, the OLS estimator’s frequentist property of consistency in combination with taking the median of a large set of ‘combined draws’  $\tilde{y}_{T+1}^s$  implies that OLS is still a usable approach. For example, the use of Generalized Least Squares (GLS) methods would not yield substantially different forecasts  $\hat{y}_{T+1}$ . The impact of this assumption on the ‘shrunk’ predictive density is arguably small; a closer look at this issue is left as a topic for further research.

### Scheme 3: Time-varying weights (TVW)

The complementary roles of different models in approximating the data generating process may differ over time. Therefore, substantially better forecasts may be obtained by extending (5.6) to allow the model weights  $w_i$  ( $i = 1, \dots, n$ ) to change over time, resulting in

$$y_t = w_{t,0} + \sum_{i=1}^n w_{t,i} y_{t,i} + u_t \quad u_t \sim N(0, \sigma^2) \quad t = 1, 2, \dots, T. \quad (5.9)$$

Terui and Van Dijk (2002) have proposed a method that extends the linear weight combination of point forecasts to time-varying weights. We extend their approach by making use of Bayesian density forecasts, taking into account parameter uncertainty. As Terui and Van Dijk (2002) we assume that the model weights  $w_t = (w_{t,0}, w_{t,1}, \dots, w_{t,n})'$  ( $t = 1, \dots, T$ ) evolve over time in the following fashion:

$$w_t = w_{t-1} + \xi_t \quad \xi_t \sim N(0, \Sigma). \quad (5.10)$$

We restrict the covariance matrix  $\Sigma$  of the ‘weight innovations’  $\xi_t$  to be a diagonal matrix. The assumed independence of the weight innovations does not rule out that a posteriori

there will be coinciding (large) changes of model weights. It means that this dependence is not imposed a priori. Including correlations in the weights would make the estimation procedure computationally more difficult, and guessing in the correlation structure can be dangerous, possibly resulting in a poor forecasting scheme. Still, we intend to analyze the extension of our scheme to non-diagonal  $\Sigma$  in future research.

As in scheme 2, our algorithm results in a set of generated model weights  $w_{T+1}^s$  ( $s = 1, \dots, S$ ) given the data  $y_{1:T}$  and draws  $y_{t,i}^s$  simulated from the predictive densities  $p(y_t|D_{t-1}, m_i)$  ( $t = 1, \dots, T$ ). The generated model weights  $w_{T+1}^s$  are used to transform draws  $y_{T+1,i}^s$  ( $i = 1, \dots, n$ ) from the predictive densities  $p(y_{T+1}|D_T, m_i)$  into ‘combined draws’  $\tilde{y}_{T+1}^s$ :

$$\tilde{y}_{T+1}^s = w_{T+1,0}^s + \sum_{i=1}^n w_{T+1,i}^s y_{T+1,i}^s \quad (5.11)$$

where the median of  $\tilde{y}_{T+1}^s$  ( $s = 1, \dots, S$ ) is our point forecast  $\hat{y}_{T+1}$  for  $y_{T+1}$ . In scheme 3, a Kalman filter algorithm (see for example Harvey (1993)) having the interpretation of a Bayesian learning approach is used to iteratively update the subsequent model weights  $w_t^s$  ( $t = 1, \dots, T+1$ ) in the model given by

$$y_t = w_{t,0}^s + \sum_{i=1}^n w_{t,i}^s y_{t,i}^s + u_t^s \quad u_t^s \sim N(0, \sigma^2) \quad t = 1, 2, \dots, T \quad (5.12)$$

and (5.10). We fix the values of  $\sigma^2$  and the diagonal elements of  $\Sigma$ . A Bayesian can interpret these assumptions as having priors on  $\sigma^2$  and  $\Sigma$  with zero variances. For each  $s$  the parameters  $\sigma^2$  and  $\Sigma$  could also be estimated by maximum likelihood or MCMC methods, but we discard this to reduce computational time.<sup>4</sup>

The model weights  $w_t^s$  incorporate a trade-off between minimizing the differences between the observed values  $y_{1:T}$  and linear combinations of ‘predicted’ values  $y_{1:T,i}^s$  ( $i = 1, \dots, n$ ), and constructing a ‘smooth’ path of weights  $w_t^s$  over time.

#### Scheme 4: Robust time-varying weights (RTVW)

Recently, a new specification has been developed that makes parameter estimation in case of instability over time more robust to prior assumptions, see for example Giordani and Villani (2008) and Groen et al. (2009) for applications. We extend the scheme 3 of time-varying model weights following the same reasoning. Then the weight innovations are equal to the latent variables  $\xi_{t,i}$  ( $i = 0, 1, \dots, n$ ) only with probability  $\pi_i$  and set equal to 0 with probability  $1 - \pi_i$ . That is, equation (5.10) becomes

$$w_t = w_{t-1} + k_t \odot \xi_t \quad \xi_t \sim N(0, \Sigma) \quad (5.13)$$

with  $k_t = (k_{0,t}, k_{1,t}, \dots, k_{n,t})'$ , where each element  $k_{i,t}$  of the vector  $k_t$  is an unobserved 0/1 variable with  $P[k_{i,t} = 1] = \pi_i$ . The Hadamard product  $\odot$  refers to element-by-element multiplication.  $\Sigma$  is again restricted to be a diagonal matrix.

<sup>4</sup>In the financial application (with  $n = 4$  models) we set  $\sigma^2$  equal to its OLS estimate in (5.6) allowing it to change with  $s$ . The  $(n+1) \times 1$  vector  $\text{diag}(\Sigma)$  of diagonal elements of  $\Sigma$  is set as  $(0.1, 0.01 \iota_n')'$  with  $\iota_n$  the  $n \times 1$  vector consisting of ones, to have (small) signal-to-noise ratios in the range from 0.01 to 0.005. For robustness we have tried different values of  $\sigma^2$  and  $\Sigma$  with signal-to-noise ratios ranging from 0.0001 to 0.1, all resulting in qualitatively equal results. In the macroeconomic application we set  $\text{diag}(\Sigma) = (0.01, 0.005 \iota_n')'$ .

The model (5.12)-(5.13) is estimated following Gerlach et al. (2000), estimating  $k_t$  by deriving its posterior density conditional on  $\sigma^2$  and  $\Sigma$ , but not on  $w_t$ . Then, we apply the Kalman Filter to estimate the latent factors  $w_t$ . We set  $\sigma^2$  and the diagonal elements of  $\Sigma$  to the same fixed values as for scheme 3.

### 5.3 Financial application

In our first application we investigate the forecasting performance and economic gains obtained by applying the four forecast combination schemes to the case of US stock index returns, the continuously compounded monthly return on the S&P 500 index in excess of the 1-month T-Bill rate, from January 1966 to December 2008, for a total of 516 observations.

We use  $n = 4$  individual models. The first model is based on the idea that a set of financial and macroeconomic variables contains potentially relevant factors for forecasting stock returns. Among others, Pesaran and Timmermann (1995), Cremers (2002), Marquering and Verbeek (2004) have shown that such variables can have predictive power. We include as predictors the S&P 500 index dividend yield defined as the ratio of dividends over the previous twelve months and the current stock price, the 3-month T-Bill rate, the monthly change in the 3-month T-bill rate, the term spread defined as the difference between the 10-year T-bond rate and the 3-month T-bill rate, the credit spread defined as the difference between Moody's Baa and Aaa yields, the yield spread defined as the difference between the Federal funds rate and the 3-month T-bill rate, the annual inflation rate based on the producer price index (PPI) for finished goods, the annual growth rate of industrial production, and the annual growth rate of the monetary base measure M1. We take into account the typical publication lag of macroeconomic variables in order to avoid look-ahead bias and we include inflation, the growth rates of industrial production and the monetary base with a two-month lag. As the financial variables are promptly available, these are included with a one-month lag. We label this forecasting model "Leading indicator" (LI).

The second forecasting model is a simple linear regression model with a constant and a dummy for November-April. It is based on the popular market saying "Sell in May and go away", also known as the "Halloween indicator" (HI) which is based on the assumption that stock returns can be predicted simply by deterministic time patterns. This suggests to buy stock in November and sell it in May. Bouman and Jacobsen (2002) show that this strategy has predictive power.

The third model allows for a well-known stylized fact on excess returns, time-varying volatility. We apply a stochastic volatility (SV) model with time varying mean:

$$r_t = \mu_t + \sigma_t u_t \quad u_t \sim N(0, 1) \quad (5.14)$$

$$\mu_t = \mu_{t-1} + \xi_{1,t} \quad \xi_{1,t} \sim N(0, \tau_1^2) \quad (5.15)$$

$$\ln(\sigma_t^2) = \ln(\sigma_{t-1}^2) + \xi_{2,t} \quad \xi_{2,t} \sim N(0, \tau_2^2) \quad (5.16)$$

The fourth model is a robust extension of the SV model that allows for parameter instability as in Giordani and Kohn (2008). In this robust stochastic volatility (RSV)

model the time-varying mean and volatility are given by

$$r_t = \mu_t + \sigma_t u_t \quad u_t \sim N(0, 1) \quad (5.17)$$

$$\mu_t = \mu_{t-1} + K_{1,t} \xi_{1,t} \quad \xi_{1,t} \sim N(0, \tau_1^2) \quad (5.18)$$

$$\ln(\sigma_t^2) = \ln(\sigma_{t-1}^2) + K_{2,t} \xi_{2,t} \quad \xi_{2,t} \sim N(0, \tau_2^2) \quad (5.19)$$

where  $K_{j,t}$  ( $j = 1, 2$ ;  $t = 1, \dots, T$ ) is an unobserved 0/1 variable with  $P[K_{j,t} = 1] = \pi_{j,RSV}$ .

The LI and HI specifications are linear models, therefore standard Bayesian derivations apply to these, see for example Koop (2003). For estimation of the SV and RSV models we refer to Giordani et al. (2007).

### 5.3.1 Evaluation

We evaluate the statistical accuracy of the individual models and the four forecast combination schemes in terms of the root mean square error (RMSPE), and in terms of the correctly predicted percentage of sign (Sign Ratio). Moreover, as an investor is more interested in the economic value of a forecasting model than its precision, we test our conclusions in an active short-term investment exercise, with an investment horizon of one month.

For this economic value criterion we closely follow the approach taken in Ravazzolo, Paap, Van Dijk, and Franses (2008), which we describe below. The investor's portfolio consists of a stock index and riskfree bonds only. At the start of each month  $T + 1$ , the investor decides upon the fraction of her portfolio to be invested in stocks  $pw_{T+1}$ , based upon a forecast of the excess stock return  $r_{T+1}$ . The investor is assumed to maximize a power utility function with coefficient of relative risk aversion  $\gamma$ :

$$u(W_{T+1}) = \frac{W_{T+1}^{1-\gamma}}{1-\gamma}, \quad \gamma > 1, \quad (5.20)$$

where  $W_{T+1}$  is the wealth at the end of period  $T + 1$ , which is equal to

$$W_{T+1} = W_T ((1 - pw_{T+1}) \exp(r_{f,T+1}) + pw_{T+1} \exp(r_{f,T+1} + r_{T+1})), \quad (5.21)$$

where  $W_T$  denotes initial wealth, and where  $r_{f,T+1}$  is the riskfree rate.

Without loss of generality we set initial wealth equal to one,  $W_T = 1$ , such that the investor's optimization problem is given by

$$\begin{aligned} \max_{pw_{T+1}} E_T(u(W_{T+1})) = \\ \max_{pw_{T+1}} E_T \left( \frac{((1 - pw_{T+1}) \exp(r_{f,T+1}) + pw_{T+1} \exp(r_{f,T+1} + r_{T+1}))^{1-\gamma}}{1-\gamma} \right), \end{aligned} \quad (5.22)$$

where  $E_T$  is the conditional expectation given information  $D_T$  at time  $T$ . How this expectation is computed depends on how the predictive density for the excess returns is computed. If we generally denote this density as  $p(r_{T+1}|D_T)$ , the investor solves the following problem:

$$\max_{pw_{T+1}} \int u(W_{T+1}) p(r_{T+1}|D_T) dr_{T+1}. \quad (5.23)$$

The integral in (5.23) is approximated by generating  $G$  independent draws  $\{r_{T+1}^g\}_{g=1}^G$  from the predictive density  $p(r_{T+1}|D_T)$ , and then using a numerical optimization method to maximize the quantity:

$$\frac{1}{G} \sum_{g=1}^G \left( \frac{((1 - pw_{T+1}) \exp(r_{f,T+1}) + pw_{T+1} \exp(r_{f,T+1} + r_{T+1}^g))^{1-\gamma}}{1 - \gamma} \right) \quad (5.24)$$

We do not allow for short-sales or leveraging, constraining  $pw_{T+1}$  to be in the  $[0,1]$  interval (see Barberis (2000)).

We include eight cases in the empirical analysis below. We consider an investor who obtains a forecast of the excess stock return  $r_{T+1}$  from the  $n = 4$  individual models (denoted LI, HI, SV and RSV) described above. Then, we consider combination forecasts using the four schemes (BMA, LIN, TVW and RTVW) from section 5.2, where all the individual models are combined.

We evaluate the different investment strategies by computing the *ex post* annualized mean portfolio return, the annualized standard deviation, the annualized Sharpe ratio and the total utility. Utility levels are computed by substituting the realized return of the portfolios at time  $T + 1$  into (5.20). Total utility is then obtained as the sum of  $u(W_{T+1})$  across all  $T^*$  investment periods  $T = T_0 + 1, \dots, T_0 + T^*$ , where the first investment decision is made at the end of period  $T_0$ . In order to compare alternative strategies we compute the multiplication factor of wealth that would equate their average utilities. For example, suppose we compare two strategies A and B. The wealth provided at time  $T + 1$  by the two resulting portfolios is denoted as  $W_{A,T+1}$  and  $W_{B,T+1}$ , respectively. We then determine the value of  $\Delta$  such that

$$\sum_{T=T_0}^{T_0+T^*-1} u(W_{A,T+1}) = \sum_{T=T_0}^{T_0+T^*-1} u(W_{B,T+1}/\exp(\Delta)). \quad (5.25)$$

Following Fleming et al. (2001), we interpret  $\Delta$  as the maximum performance fee the investor would be willing to pay to switch from strategy A to strategy B. For comparison of multiple investment strategies, it is useful to note that – under a power utility specification – the performance fee an investor is willing to pay to switch from strategy A to strategy B can also be computed as the difference between the performance fees of these strategies with respect to a third strategy C.<sup>5</sup> We use this property below to infer the added value of strategies based on individual models and combination schemes by computing  $\Delta$  with respect to three static benchmark strategies: holding stocks only ( $\Delta_s$ ), holding a portfolio consisting of 50% stocks and 50% bonds ( $\Delta_m$ ), and holding bonds only ( $\Delta_b$ ).

Finally, the portfolio weights in the active investment strategies change every month, and the portfolio must be rebalanced accordingly. Hence, transaction costs play a non-trivial role and should be taken into account when evaluating the relative performance of different strategies. Rebalancing the portfolio at the start of month  $T + 1$  means that

<sup>5</sup>This follows from the fact that combining (5.25) for the comparisons of strategies A and B with C,  $\sum_T u(W_{C,T+1}) = \sum_T u(W_{A,T+1}/\exp(\Delta_A))$  and  $\sum_T u(W_{C,T+1}) = \sum_T u(W_{B,T+1}/\exp(\Delta_B))$ , gives  $\sum_T u(W_{A,T+1}/\exp(\Delta_A)) = \sum_T u(W_{B,T+1}/\exp(\Delta_B))$ . Using the power utility specification in (5.20), this can be rewritten as  $\sum_T u(W_{A,T+1}) = \sum_T u(W_{B,T+1}/\exp(\Delta_B - \Delta_A))$ .



the weight invested in stocks is changed from  $pw_T$  to  $pw_{T+1}$ . We assume that transaction costs amount to a fixed percentage  $c$  on each traded dollar. Setting the initial wealth  $W_T$  equal to 1 for simplicity, transaction costs at time  $T + 1$  are equal to

$$c_{T+1} = 2c|pw_{T+1} - pw_T| \quad (5.26)$$

where the multiplication by 2 follows from the fact that the investor rebalances her investments in both stocks and bonds. The net excess portfolio return is then given by  $r_{T+1} - c_{T+1}$ . We apply a scenario with transaction costs of 0.1%.

### 5.3.2 Empirical Results

The analysis for the active investment strategies is implemented for the period from January 1987 until December 2008, involving  $T^* = 264$  one month ahead excess stock return forecasts. The individual models are estimated recursively using an expanding window of observations. The initial 12 predictions for each individual model are used as training period for combination schemes and making the first combined prediction. The investment strategies are implemented for a level of relative risk aversion of  $\gamma = 6$ .<sup>6</sup>

Before we analyze the performance of the different portfolios, we summarize the statistical accuracy of the excess return forecasts. All the individual models give similar RMSPE statistics in Table 5.1, for the RSV model just the smallest and for the LI model the highest. The sign ratio is the highest for the SV model, but hardly exceeds 60%, indicating low predictability. Due to this low predictability, small differences in RMSPE may have substantial economic value. We investigate this in the portfolio exercise. The SV model gives the highest Sharpe ratio, realized final utility and comparison fees  $\Delta$  among the individual models. The TVW and RTVW combination schemes, however, provide much higher statistics; in particular RTVW outperforms all the other models in terms of Sharpe ratio and realized utility value, and all three  $\Delta$ 's are positive. Figure 5.1 can help to explain these findings. Individual models allocate too low weight to the risky asset resulting in low portfolio returns. BMA has a similar problem. The LIN, TVW and RTVW combinations allocate higher weights to the stock asset, but RTVW is the only scheme that drastically reduces this weight in bear market periods as the burst of the internet bubble in 2001-2003 or the recent financial crisis in the second part of 2007 and 2008. Panel C in Table 5.1 shows evidence that the findings are similar when taking into account the presence of medium transaction costs.

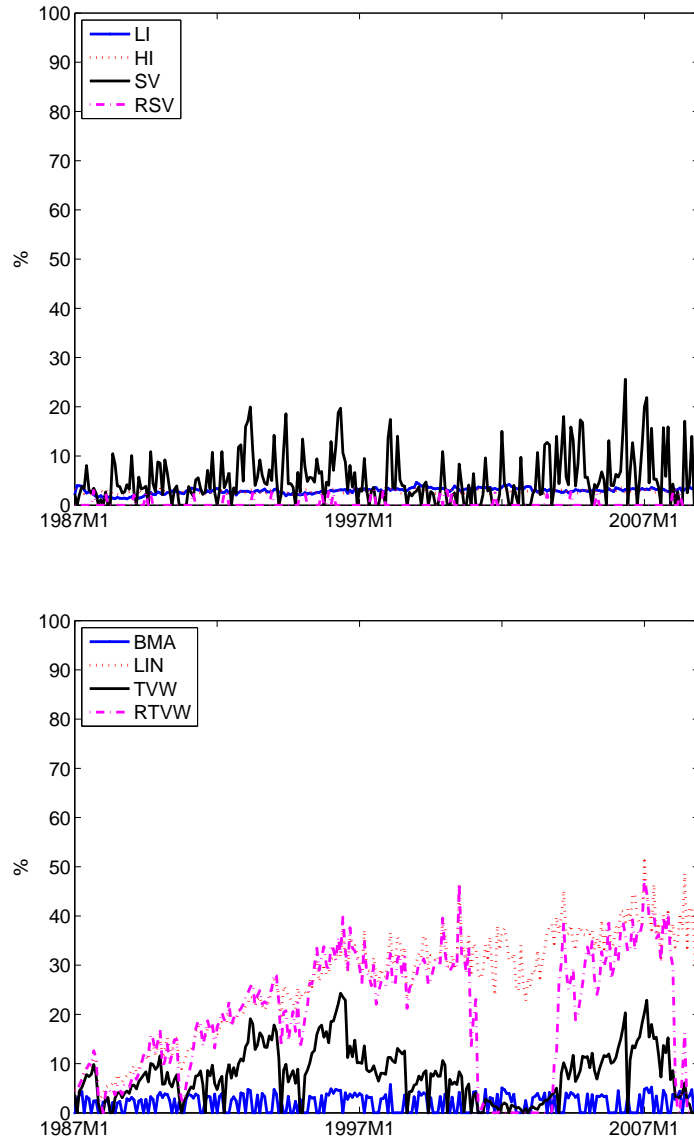
The good performance of RTVW as compared to LIN and TVW shows that its robust flexible structure pays off. The higher portfolio weight of stock in bull markets for RTVW, as compared to the individual models and BMA, is due to the 'shrunk' predictive density. This 'shrunk' excess return distribution is not so much 'compressed' that the risky asset's portfolio weight switches from 0% to 100% when its mean changes from negative to positive values. Rather, the parameter and model uncertainty that are incorporated in this 'shrunk' predictive density imply an investment strategy with a

<sup>6</sup>We also implement exercises with  $\gamma = 4$  and  $\gamma = 8$ . Results are qualitatively similar and available upon request.

Table 5.1: Financial application: statistical and economic performance

	LI	HI	SV	RSV	BMA	LIN	TVW	RTVW
<b>Panel A: Statistical accuracy</b>								
RMSPE	4.618	4.478	4.509	4.470	4.500	4.514	4.484	4.485
Sign Ratio	0.527	0.549	0.614	0.598	0.587	0.610	0.602	0.598
<b>Panel B: Active portfolio performances, <math>\gamma = 6</math>, transaction costs <math>c = 0</math></b>								
Portfolio mean	4.708	4.741	4.812	4.657	4.701	5.177	5.021	5.785
Portfolio st dev	0.794	0.769	1.139	0.614	0.739	4.356	1.332	3.062
Sharpe ratio	0.110	0.156	0.168	0.060	0.108	0.128	0.301	0.380
Realized Utility	-51.77	-51.76	-51.75	-51.79	-51.77	-51.73	-51.70	-51.56
$\Delta_s$	285.5	288.7	295.2	277.9	283.8	304.3	317.1	381.3
$\Delta_m$	-63.71	-60.49	-54.03	-71.29	-65.42	-44.95	-32.10	32.10
$\Delta_b$	11.46	14.68	21.14	3.876	9.748	30.22	43.07	107.3
<b>Panel C: Active portfolio performances, <math>\gamma = 6</math>, transaction costs <math>c=10</math> bp</b>								
Portfolio return	4.708	4.740	4.811	4.657	4.700	5.176	5.020	5.784
Portfolio st dev	0.794	0.769	1.139	0.614	0.739	4.355	1.332	3.062
Sharpe ratio	0.110	0.156	0.167	0.060	0.108	0.128	0.300	0.380
Realized Utility	-51.77	-51.77	-51.77	-51.79	-51.78	-51.75	-51.71	-51.58
$\Delta_s$	284.7	287.9	284.5	276.6	279.1	297.1	311.7	373.6
$\Delta_m$	-64.65	-61.42	-64.80	-72.72	-70.18	-52.18	-37.66	24.31
$\Delta_b$	10.81	14.04	10.67	2.741	5.289	23.29	37.81	99.77

Figure 5.1: Financial application: portfolio weight of stock (S&amp;P500)



*Note:* The graphs show the portfolio weight on the risky asset (S&P500) over the out-of-sample period associated to active asset management given by individual models in the top panel and combination schemes in the bottom panel.

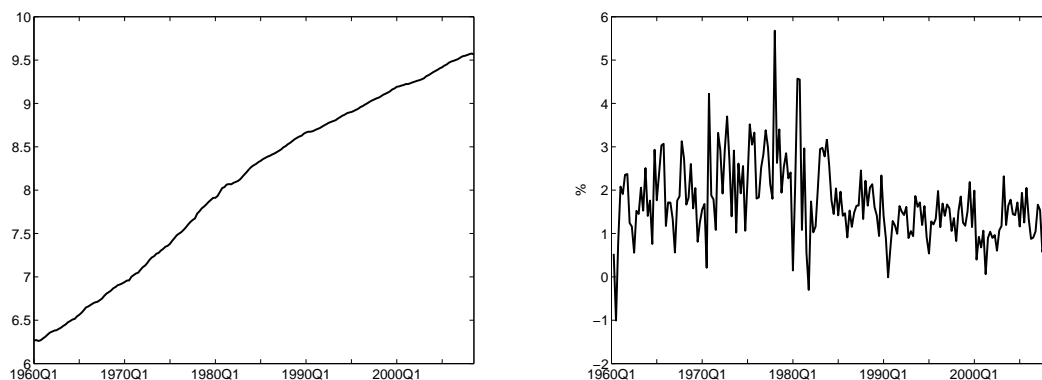
smooth, ‘moderate’, yet flexible evolvement of the risky asset’s portfolio weight over time. Lettau and Van Nieuwerburgh (2008) find that the uncertainty on the size of *steady-state* shifts rather than their dates is responsible for the difficulty of forecasting stock returns in real time. The ‘shrunk’ predictive density of the RTVW scheme may be particularly informative on the current and future evolvement of this steady-state, the driving force of return predictability. This may be the explanation for the RTVW scheme’s good results.

We intend to analyze its performance in other portfolio management exercises in future research, in order to investigate the robustness of our findings.

## 5.4 US real GDP Growth

We now perform an empirical analysis on a key macroeconomic series, the U.S. real Gross Domestic Product (GDP) growth. We collected real GDP (seasonally adjusted) figures from the U.S. Department of Commerce, Bureau of Economic Analysis. The left panel of Figure 5.2 plots the log quarterly GDP level for our sample 1960:Q1 to 2008:Q3 (195 observations) and shows that GDP has followed an upward sloping pattern but with fluctuations around this trend. The quarterly growth rate,  $\ln GDP_t - \ln GDP_{t-1}$ , shown in the right panel of Figure 5.2, underlines these fluctuations with periods of positive changes followed by periods of negative changes, clearly indicating business cycles; for more details we refer to Harvey et al. (2007).

Figure 5.2: US real GDP



*Note:* Quarterly log levels of U.S. real GDP (left) and quarterly GDP growth rate in % terms (right). The sample is 1960:Q1 - 2008:Q3.

As in the previous section, we apply various linear and nonlinear models and forecast combinations to assess these models' suitability in a pseudo-real-time out-of-sample forecasting exercise. In the forecast exercise we use an initial in-sample period from 1960:Q1 to 1979:Q4 to obtain initial parameter estimates and we forecast the GDP growth figure for 1980:Q1. We then expand the estimation sample with the value in 1980:Q1, re-estimating the parameters, and we forecast the next value for 1980:Q2. We continue this procedure up to the last value and we end up with a total of 115 forecasts.

We apply  $n = 6$  individual time series models to infer and forecast GDP. Four models are linear specifications, two models are time-varying parameter specifications. The first and second model are random walk models, without and with drift (RW and RWD). The third model is the autoregressive (AR) model of order 1. We follow Schotman and Van Dijk (1991) and specify a weakly informative 'regularization' prior that helps to prevent problems that could be encountered during the estimation using the Gibbs sampler, if a

flat prior were used. The fourth model we apply is an error correction model (ECM). We apply the same model as in De Pooter et al. (2008):

$$\Delta y_t = \delta + (\rho_1 + \rho_2 - 1)(y_{t-1} - \mu - \delta(t-1)) - \rho_2(\Delta y_{t-1} - \delta) + \varepsilon_t, \quad \varepsilon_t \sim N(0, \sigma^2), \quad (5.27)$$

which can be rewritten as:

$$y_t - \delta t = (1 - \rho_1 - \rho_2)\mu + \rho_1(y_{t-1} - \delta(t-1)) + \rho_2(y_{t-2} - \delta(t-2)) + \varepsilon_t, \quad \varepsilon_t \sim N(0, \sigma^2). \quad (5.28)$$

The prior that we use is an extension of the prior of Schotman and Van Dijk (1991). The fifth and sixth models are a state-space model (SSM) and its robust extension (RSSM), that are given by the SV and RSV models of section 5.3.

We use the root mean square prediction error (RMSPE) to compare different point forecasts. Table 5.2 shows that the random walk models perform poorly. For all other models, the test of Clark and West (2007) for equal forecasting quality of nested models rejects the null hypothesis versus the RW model. The AR model is a bit more precise than the ECM. The models with time varying parameters, SSM and RSSM, perform very well. Figure 5.3 shows that all models with fixed parameters perform poorly when GDP decreases rapidly and substantially as in NBER recessions, and it takes some quarters for models to adjust, in particular in the 2001 recession and the 2008 recession. Time-varying parameter models seem to cope better with this.

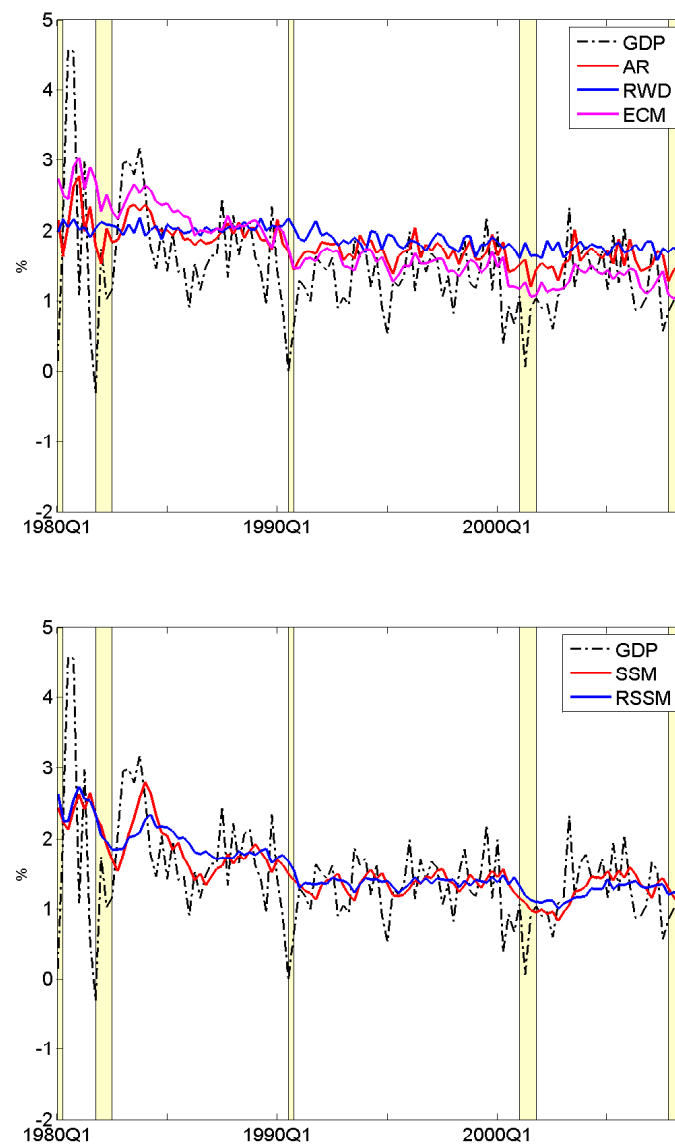
Table 5.2: Forecasting U.S. real GDP growth (in % terms): root mean square prediction error (RMSPE)

individual models		combinations	
RW	1.650	BMA	0.718
RWD	0.863	LIN	0.829
AR	0.772	TVW	0.757
ECM	0.790	RTVW	0.727
SSM	0.730		
RSSM	0.747		

The BMA and RTVW combination schemes provide even better statistics than the SSM and RSSM models. LIN is the worst averaging scheme; LIN performs similarly to the AR and ECM models. Figure 5.4 shows that LIN is performing particularly poorly in the 1980's and 1990's. Weight estimates for this scheme may be highly inaccurate as the number of individual models is relatively large and instability possibly high. Moreover, Figure 5.4 indicates that the other averaging schemes react much faster to sharp decreases in GDP. Especially the RTVW scheme may early indicate recessions: before both the 1991 and 2001 crises its point forecast decreases substantially with approximately 0.5%.

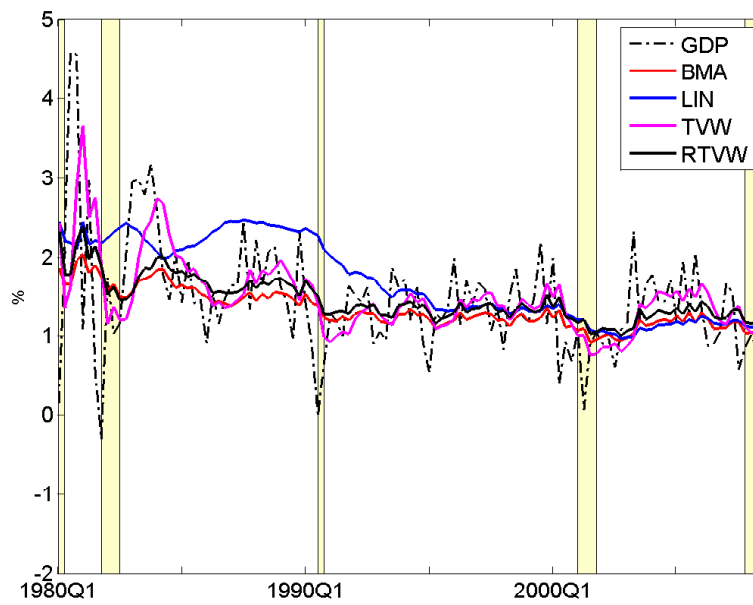
To sum up, our results suggest that model averaging may be very beneficial in business cycle analysis and forecasting. The combination method must, however, be chosen carefully and it should cope with estimation efficiency and structural instability, in particular if weights are estimated in regression equations. Again, more extensive studies should be performed to investigate the robustness of our findings, for example over different countries and periods.

Figure 5.3: US real GDP: point forecasts



*Note:* Quarterly U.S. real GDP growth (in %) and point forecasts given by individual models. Vertical bars highlight NBER recession periods.

Figure 5.4: US real GDP: point forecasts



*Note:* Quarterly U.S. real GDP growth (in %) and point forecasts given by combination schemes. Vertical bars highlight NBER recession periods.

## 5.5 Final remarks

The empirical applications have indicated, firstly, that averaging strategies can give higher predictive quality than selecting the best model; secondly, that properly specified time varying model weights yield higher forecast accuracy and substantial economic gains compared with other averaging schemes. The presented results lead to multiple directions for future research. As we already mentioned, interesting possibilities for further research are a rigorous analysis of the impact of some assumptions – both on theoretical aspects and practical applications – and an extensive study on the robustness of our findings.

Another topic for further research is to compare our results to other time varying weight combination schemes, such as regime switching, see e.g. Guidolin and Timmermann (2009), or schemes that carefully model breaks, see e.g. Ravazzolo et al. (2008). For the application to portfolio management, a natural extension is the prediction of multivariate returns processes. The proposed combination schemes can also be adapted to the specific prediction of variance, skewness or kurtosis.



# Nederlandse Samenvatting (Summary in Dutch)

Zowel binnen de Bayesiaanse als de frequentistische econometrische benadering kunnen verschillende modellen tot verschillende uitkomsten leiden. Een veelgebruikte aanpak in een dergelijke situatie is het toepassen van modelselectie: één enkel model wordt geselecteerd uit een veelheid van mogelijke modellen. Analyses, zoals bijvoorbeeld het genereren van voorspellingen, worden vervolgens gebaseerd op dit geselecteerde model.

Mogelijk gevolg is dat bij modelselectie de informatie die de modellen, die niet geselecteerd zijn, zouden kunnen opleveren verder buiten beschouwing gelaten wordt. Om die reden beschouwen we in dit proefschrift behalve modelselectie ook modelmiddeling, waarbij informatie vanuit meerdere modellen wordt gecombineerd. Het idee hierachter is dat één enkel model niet onder alle omstandigheden het beste model zal zijn.

Modelmiddeling is een manier om resultaten uit meerdere modellen te combineren, zodanig dat met informatie uit al deze modellen rekening gehouden wordt bijvoorbeeld voor het bepalen van voorspellingen. De Bayesiaanse benadering biedt een natuurlijk kader om om te gaan met onzekerheid en is in het bijzonder ook geschikt voor het omgaan met modelonzekerheid door middel van modelmiddeling. In de Hoofdstukken 4 en 5 van dit proefschrift wordt Bayesiaanse modelmiddeling toegepast op dynamische econometrische modellen voor tijdreeksen. In het bijzonder ontwikkelen we methoden waarin het gewicht van de modellen over de tijd kan variëren en onderzoeken de toegevoegde waarde van deze extra flexibiliteit. Enkele van de econometrische methoden en technieken die in de Hoofdstukken 2 en 3 gebruikt en ontwikkeld worden, dienen ook als bouwsteen voor de latere hoofdstukken.

Dit proefschrift behandelt enkele thema's uit de Bayesiaanse econometrie. In de Bayesiaanse benadering kan een modelgewicht beschouwd worden als een kans. We hanteren waar mogelijk het vertrekpunt dat de data zoveel mogelijk bepalend zijn voor de uitkomsten van het wetenschappelijke model. Om deze reden specificeren we in dit proefschrift voor de parameters van het model in de meeste gevallen niet-informatieve priorverdelingen. Een niet-informatieve priorverdeling weerspiegelt dat er geen andere informatie gebruikt wordt dan die uit de combinatie van de waargenomen data en het veronderstelde model kan worden afgeleid.

In Hoofdstuk 2 onderzoeken we de implicaties van het specificeren van een niet-informatieve priorverdeling voor een cointegratiemodel. In het bijzonder onderzoeken we de voorwaarden waaronder posterior kansen en hogere momenten van de kansvariabelen bestaan. We ontwikkelen een alternatieve parametrisatie en normalisatie van de parameters van het model, waarin alle parameters getransformeerd worden naar een begrensde ruimte, zodat het mogelijk wordt een integreerbare uniforme, niet-informatieve priorverdeling te specificeren.

In Hoofdstuk 3 introduceren we een op simulatiemethoden gebaseerde Bayesiaanse analyse van een model met latente, niet-geobserveerde kansvariabelen die het mogelijk maakt het onderliggende gemiddelde van reële wisselkoersen zodanig te modelleren dat deze niet constant is, maar langzaam kan variëren door de tijd. Hoewel het achterliggende idee aansluit op Engel (2000), hanteren we een andere specificatie van het model om in het frequentiedomein scherper onderscheid te kunnen maken tussen de trendcomponent en de stationaire componenten. In de analyse van het model maken we gebruik van zogenaamde *Markov Chain Monte Carlo* simulatiemethoden, zoals *Gibbs*

*sampling* en *Metropolis-Hastings* in combinatie met de techniek van data-augmentatie, die het mogelijk maakt om met de niet-geobserveerde kansvariabelen om te gaan. Ten behoeve van de *Metropolis-Hastings*-stap in de simulatie ontwikkelen we een zorgvuldig gekozen kandidaat-kansverdeling, die het mogelijk maakt met grote efficiëntie uit een niet-standaard kansverdeling te simuleren. De simulatietechniek maakt het mogelijk om de volledige kansverdeling van complexe functies van de parameters van het model te bepalen, in het bijzonder wordt dit toegepast op een maatstaf die aangeeft hoe de invloed van een verstoring door de tijd heen doorwerkt en op de halfwaardetijd, dat is de tijd die verstrijkt voor de invloed van een verstoring voor minimaal de helft is uitgedoofd. Met dit model worden modelkansen berekend. De data geven weinig ondersteuning voor stationariteit van de reële wisselkoersen. De hypothese van koopkrachtpariteit op lange termijn wordt daarmee dus niet door de data ondersteund. De kortere termijn terugkeer naar het tijdsvariërende gemiddelde blijkt echter wel dichtbij economisch plausible waarden te liggen en geven daarmee wel een indicatie voor het bestaan van een zwakkere vorm van koopkrachtpariteit op korte termijn. De meerwaarde van de gehanteerde specificatie van het model is dat de componenten die de korte termijndynamiek beschrijven behulpzaam kunnen zijn bij het modelleren en voorspellen van reële wisselkoersen.

In Hoofdstuk 4 introduceren we een nieuwe klasse van tijdreeksmodellen, die nauw verwant is aan de Fourier decompositie van tijdreeksen in cyclische componenten. De modellen worden gekenmerkt door een flexibel aantal stochastische cyclische componenten. We analyseren deze modelklasse met behulp van een simulatie-algoritme en bepalen daarmee ook modelkansen. Op basis hiervan toetsen we enerzijds op het meest waarschijnlijke aantal cyclische componenten. Anderzijds dienen deze modelkansen ook als modelgewichten in de toepassing van Bayesiaanse modelmiddeling. Hiermee wordt het mogelijk een kansverdeling voor voorspellingen te maken die rekening houdt met onzekerheid voortkomend uit vijf verschillende bronnen, namelijk de onzekerheid over het toekomstige verstoringsproces, onzekerheid over de exacte waarde van de parameters van het model, onzekerheid over de niet-waargenomen componenten, modelonzekerheid over het aantal cyclische componenten en modelonzekerheid over de stationariteit van het proces.

Hoofdstuk 5 breidt Bayesiaanse modelmiddeling uit met drie verschillende methoden om voorspellingen van verschillende modellen te combineren. Deze methoden houden niet alleen rekening met zowel parameteronzekerheid als modelonzekerheid, maar voegen daaraan ook de mogelijkheid toe dat de modelgewichten door de tijd kunnen variëren. Deze technieken worden gebruikt in een tweetal empirische toepassingen, waarbij de verdiensten van de methoden beoordeeld worden. Onze bevindingen geven aan dat modelmiddeling tot een hogere voorspelkwaliteit leidt dan het gebruik van het selecteren van één model. In het bijzonder blijkt de robuuste tijdsvariërende methode niet alleen de hoogste voorspelnauwkeurigheid op te leveren, maar ook substantiële voordelen in economische termen.



# Bibliography

- T.W. Anderson. *An introduction to multivariate statistical analysis*. Wiley, New York, 3rd edition, 2003.
- D. Ardia, L.F. Hoogerheide, and H.K. Van Dijk. To bridge, to warp or to wrap? A comparative study of Monte Carlo methods for efficient evaluation of marginal likelihoods. Tinbergen institute report 09-017/4, 2009.
- N. Barberis. Investing for the long run when returns are predictable. *Journal of Finance*, 55:225–264, 2000.
- N. Baştürk, C. Çakmaklı, S. P. Ceyhan, and H.K. Van Dijk. Posterior-predictive evidence on US inflation using extended Phillips Curve models with non-filtered data. *Journal of Applied Econometrics*, 29(7):1164–1182, 2014a. ISSN 1099-1255. doi: 10.1002/jae.2411. URL <http://dx.doi.org/10.1002/jae.2411>.
- N. Baştürk, C. Çakmaklı, S.P. Ceyhan, and H.K. Van Dijk. On the rise of Bayesian econometrics after Cowles Foundation monographs 10, 14. *Æconomia*, 4:381–447, 2014b.
- N. Baştürk, L.F. Hoogerheide, R.H. Kleijn, and H.K. Van Dijk. Prior ignorance, likelihood shape and posterior existence in a cointegration model. unpublished manuscript, November 2015.
- J. M. Bates and C. W. J. Granger. Combination of forecasts. *Operational Research Quarterly*, 20(4):451–468, 1969.
- L. Bauwens, M. Lubrano, and J.-F. Richard. *Bayesian inference in dynamic econometric models*. Advanced texts in Econometrics. Oxford University Press, Oxford, 1999.
- S. Beveridge and C.R. Nelson. 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, 1981.
- M. Billio, R. Casarin, F. Ravazzolo, and H.K. Van Dijk. Time-varying combinations of predictive densities using nonlinear filtering. *Journal of Econometrics*, 177(2):213–232, 2013.
- S. Bouman and B. Jacobsen. The Halloween indicator, ‘sell in May and go away’: Another puzzle. *American Economic Review*, 92 (5):1618–1635, 2002.

- C.K. Carter and R. Kohn. On Gibbs sampling for state space models. *Biometrika*, **81**: 541–553, 1994.
- R. Casarin, S. Grassi, F. Ravazzolo, and H.K. Van Dijk. Dynamic predictive density combinations for large data sets in economics and finance. Technical Report 2015-084/III, Tinbergen Institute, 2015.
- G. Casella and E.I. George. Explaining the Gibbs sampler. *The American Statistician*, **46**:167–174, 1992.
- Y-W. Cheung and S.L. Lai. On the purchasing power parity puzzle. *Journal of International Economics*, **52**:321–330, 2000.
- S. Chib. Marginal likelihood from the Gibbs output. *Journal of the American Statistical Association*, **90**(432):1313–1321, 1995.
- S. Chib and E. Greenberg. Understanding the Metropolis-Hastings algorithm. *The American Statistician*, **49**:327–335, 1995.
- S. Chib and I. Jeliaskov. Marginal likelihood from the Metropolis-Hastings output. *Journal of the American Statistical Association*, **96**(453):270–281, 2001.
- P.K. Clark. Nearly redundant parameters and measures of persistence in economic time series. *Journal of Economic Dynamics and Control*, **12**:447–461, 1988.
- T.E. Clark and K.D. West. Approximately normal tests for equal predictive accuracy in nested models. *Journal of Econometrics*, **138**(1):291–311, May 2007.
- T. Cogley, S. Morozov, and T.J. Sargent. Bayesian fan charts for U.K. inflation in an evolving monetary system. *unpublished manuscript*, 2003.
- K.J.M. Cremers. Stock return predictability: A Bayesian model selection perspective. *Review of Financial Studies*, **15**:1223–1249, 2002.
- P. de Jong and N. Shephard. The simulation smoother for time series models. *Biometrika*, **82**:339–350, 1995.
- M. De Pooter, F. Ravazzolo, R. Segers, and H.K. Van Dijk. Bayesian near-boundary analysis in basic macroeconomic time-series models. *Advances in Econometrics*, **23**: 331–402, 2008.
- J.M. Dickey. Matricvariate generalizations of the multivariate t distribution and the inverted multivariate t distribution. *The Annals of Mathematical Statistics*, **38**(2):511–518, 1967.
- T. Doan, R.B. Litterman, and C. Sims. Forecasting and Conditional Projections using Realistic Prior Distributions. *Econometric Reviews*, **3**:1–100, 1992.
- J.A. Doornik. *Object-Oriented Matrix Programming Using Ox*. Timberlake Consultants Press, London, 3rd edition, 1999. Oxford: [www.nuff.ox.ac.uk/Users/Doornik](http://www.nuff.ox.ac.uk/Users/Doornik).

- D. Draper. Assessment and propagation of model uncertainty. *Journal of the Royal Statistical Society Series B*, 56:45–98, 1995.
- C. Engel. Long-run PPP may not hold after all. *Journal of International Economics*, 57: 243–273, 2000.
- C. Fernández, E. Ley, and M.F.J. Steel. Model uncertainty in cross-country growth regressions. *Journal of Applied Econometrics*, 16:563–576, 2001.
- J. Fleming, C. Kirby, and B. Ostdiek. The economic value of volatility timing. *Journal of Finance*, 56:329–352, 2001.
- S. Frühwirth-Schnatter. Data augmentation and dynamic linear models. *Journal of Time Series Analysis*, 15:183–202, 1994.
- S. Frühwirth-Schnatter. Markov chain Monte Carlo estimation of classical and dynamic switching and mixture models. *Journal of the American Statistical Association*, 96: 194–209, 2001.
- A.E. Gelfand and A.F.M. Smith. Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85:389–409, 1990.
- R. Gerlach, C. Carter, and R. Kohn. Efficient Bayesian inference for dynamic mixture models. *Journal of the American Statistical Association*, 95:819–828, 2000.
- J. Geweke. Using simulation methods for Bayesian econometric models: inference, development and communication. *Econometric Reviews*, 18(1):1–73, 1999.
- J. Geweke and C. Whiteman. Bayesian forecasting. In G. Elliot, C.W.J. Granger, and A.G. Timmermann, editors, *Handbook of Economic Forecasting*. North-Holland, 2006.
- D. Giannone, M. Lenza, and G.E. Primiceri. Prior selection for vector autoregressions. *The Review of Economics and Statistics*, 97(2):436–451, 2015. doi: 10.1162/REST\_a.00483.
- P. Giordani and R. Kohn. Efficient Bayesian Inference for Multiple Change-Point and Mixture Innovation Models. *Journal of Business & Economic Statistics*, 26:66–77, 2008.
- P. Giordani and M. Villani. Forecasting macroeconomic time series with locally adaptive signal extraction. *Working Paper*, 2008.
- P. Giordani, R. Kohn, and D. Van Dijk. A unified approach to nonlinearity, outliers and structural breaks. *Journal of Econometrics*, 137:112–137, 2007.
- C.W.J. Granger and R. Ramanathan. Improved methods of combining forecasts. *Journal of Forecasting*, 3:197–204, 1984.
- J. Groen, R. Paap, and F. Ravazzolo. Real-time inflation forecasting in a changing world. Working paper, 2009.

- M. Guidolin and A. Timmermann. Forecasts of US Short-term Interest Rates: A Flexible Forecast Combination Approach. *Journal of Econometrics*, **150**(2):297–311, 2009.
- J.D. Hamilton. A New Approach to the Economic Analysis of Nonstationary Time Series and the Business Cycle. *Econometrica*, 57:357–384, 1989.
- J.D. Hamilton. *Time Series Analysis*. Princeton University Press, Princeton, NJ, 1994.
- J.D. Hamilton, D.F. Waggoner, and T. Zha. Normalization in econometrics. *Econometric Reviews*, 26(2–4):221–252, 2007.
- B.E. Hansen. Least squares model averaging. *Econometrica*, 75(4):1175–1189, 2007.
- D. Harding and A. Pagan. Dissecting the cycle: a methodological investigation. *Journal of Monetary Economics*, **49**:365–381, 2002.
- A.C. Harvey. *Forecasting, structural time series models and the Kalman filter*. Cambridge University Press, Cambridge, 1989.
- A.C. Harvey. *Time series models*. Harvester Wheatsheaf, 2nd edition, 1993.
- A.C. Harvey and A. Jaeger. Detrending, stylized facts and the business cycle. *Journal of Applied Econometrics*, **8**:231–247, 1993.
- A.C. Harvey and M. Streibel. Testing for deterministic versus indeterministic cycles. *Journal of Time Series Analysis*, **19**:505–529, 1998.
- A.C. Harvey, T.M. Trimbur, and H.K. Van Dijk. Cyclical components in economic time series: a Bayesian approach. Econometric Institute Report EI2002-20, Erasmus University Rotterdam, 2002.
- A.C. Harvey, T.M. Trimbur, and H.K. Van Dijk. Bayes Estimates of the Cyclical Component in Twentieth Century U.S. Gross Domestic Product. In G. L. Mazzi and G. Savio, editors, *Growth and Cycle in the Eurozone*, pages 76–89. Palgrave MacMillan, New York, 2007.
- D.F. Hendry and M.P. Clements. Pooling of forecasts. *Econometric Reviews*, 22:47–79, 2004.
- C. Hildreth. Bayesian statisticians and remote clients. *Econometrica: Journal of the Econometric Society*, 31(3):422–438, 1963.
- J. Hodges. Uncertainty, policy analysis and statistics. *Statistical Science*, 2:259–291, 1987.
- L. Hoogerheide, R.H. Kleijn, F. Ravazzolo, H.K. van Dijk, and M. Verbeek. Forecast accuracy and economic gains from Bayesian model averaging using time varying weights. *Journal of Forecasting*, **29**:251–269, 2010.
- G. Huerta and M. West. Priors and component structures in autoregressive time series models. *Journal of the Royal Statistical Society B*, **61**:881–899, 1999.



- S. Johansen. Statistical analysis of cointegration vectors. *Journal of Economic Dynamics and Control*, 12(2):231–254, 1988.
- S. Johansen. *Likelihood-based Inference in Cointegrated Vector Autoregressive Models*. Oxford University Press, 1995.
- C. Kascha and F. Ravazzolo. Combining inflation density forecasts. Norges Bank working paper 2008-22, 2008.
- R.E. Kass, L. Tierney, and A.E. Raftery. The validity of posterior expansions based on Laplace’s method. In S. Geisser, editor, *Bayesian and Likelihood methods in statistics and econometrics: Essays in honor of George A. Barnard*, pages 473–488. North Holland, Amsterdam, 1990.
- Y. Kawasaki and P.H.B.F. Franses. Do seasonal unit roots matter for forecasting monthly industrial production? *Journal of Forecasting*, **23**:77–88, 2004.
- G. Kitagawa and W. Gersch. *Smoothness Priors Analysis of Time Series*. Springer-Verlag, 1996.
- F.R. Kleibergen and R. Paap. Priors, posteriors and Bayes factors for a Bayesian analysis of cointegration. *Journal of Econometrics*, 111(2):223–249, 2002.
- F.R. Kleibergen and H.K. Van Dijk. On the shape of the likelihood/posterior in cointegration models. *Econometric Theory*, 10(3-4):514–551, 1994.
- F.R. Kleibergen and H.K. Van Dijk. Bayesian simultaneous equations analysis using reduced rank structures. *Econometric Theory*, 14(06):701–743, 1998.
- R.H. Kleijn and H.K. Van Dijk. A Bayesian analysis of the PPP puzzle using an unobserved components model. Econometric Institute Report EI2001-35, Erasmus University Rotterdam, 2001.
- R.H. Kleijn and H.K. Van Dijk. Bayes model averaging of cyclical decompositions in economic time series. *Journal of Applied Econometrics*, **21**(2):191–212, 2006.
- G. Koop. *Bayesian Econometrics*. John Wiley & Sons Ltd, West Sussex, England, 2003.
- G. Koop and H.K. Van Dijk. Testing for integration using evolving trend and seasonals models: A Bayesian Approach. *Journal of Econometrics*, **97**:261–291, 2000.
- G.M. Koop, R.W. Strachan, H.K. Van Dijk, and M. Villani. Bayesian approaches to cointegration. In *The Palgrave Handbook of Theoretical Econometrics*, pages 871–898. Palgrave Macmillan, 2006.
- S.J. Koopman, N. Shephard, and J.A. Doornik. Statistical algorithms for models in state space using SsfPack 2.2. *Econometrics Journal*, **1**:1–55, 1998.
- D. Kwiatkowski, P.C.B. Phillips, P. Schmidt, and Y. Shin. Testing the null hypothesis of stationarity against the alternative of a unit root. *Journal of Econometrics*, **54**: 159–178, 1992.

- E. Leamer. *Specification Searches*. New York: Wiley, 1978.
- M. Lettau and S. Van Nieuwerburgh. Reconciling the return predictability evidence. *The Review of Financial Studies*, 21:1607–1652, 2008.
- D.M. Madigan and A.E. Raftery. Model selection and accounting for model uncertainty in graphical models using Occam’s window. *Journal of the American Statistical Association*, 89:1335–1346, 1994.
- J.R. Magnus and H. Neudecker. *Matrix differential calculus with applications in statistics and econometrics*. John Wiley & Sons, 1995.
- M. Marcellino. Forecasting pooling for short time series of macroeconomic variables. *Oxford Bulletin of Economic and Statistics*, 66:91–112, 2004.
- W. Marquering and M. Verbeek. The economic value of predicting stock index returns and volatility. *Journal of Financial and Quantitative Analysis*, 39 (2):407–429, 2004.
- C. Min and A. Zellner. Bayesian and non-bayesian methods for combining models and forecasts with applications to forecasting international growth rates. *Journal of Econometrics*, 56:89–118, 1993.
- J. Mitchell and S.G. Hall. Evaluating, comparing and combining density forecasts using the KLIC with an application to the Bank of England and NIESR ‘fan’ charts of inflation. *Oxford Bulletin of Economics and Statistics*, **67**(s1):995–1033, 2005.
- J.C. Morley, C.R. Nelson, and E. Zivot. Why are the Beveridge-Nelson and Unobserved-Components decompositions of GDP so different? *Review of Economics and Statistics*, **85**(2):235–243, 2003.
- R.J. Muirhead. *Aspects of Multivariate Statistical Theory*. John Wiley, New York, 1982.
- R. Paap and H.K. Van Dijk. Bayes estimates of markov trends in possibly cointegrated series: An application to us consumption and income. *Journal of Business and Economic Statistics*, **21**:547–563, 2003.
- M.H. Pesaran and A. Timmermann. Predictability of stock returns: Robustness and economic significance. *Journal of Finance*, 50:1201–1228, 1995.
- M.H. Pesaran and A. Timmermann. Market timing and return predictability under model instability. *Journal of Empirical Finance*, 9:495–510, 2002.
- P.C.B. Phillips. To criticize the critics: An objective Bayesian analysis of stochastic trends. *Journal of Applied Econometrics*, 6:333–364, 1991.
- C. Planas, A. Rossi, and G. Fiorentini. The marginal likelihood of structural time series models, with application to the euroarea and us nairu. *Working Paper Series 21-08, Rimini Centre for Economic Analysis*, 2008.

- F. Ravazzolo, R. Paap, D. Van Dijk, and P.H.B.F. Franses. Bayesian model averaging in the presence of structural breaks. In M. Wohar and D. Rapach, editors, *Forecasting in the Presence of Structural Breaks and Model Uncertainty*, volume 3 of *Frontiers of economics and globalization*, chapter 15, pages 561–594. Emerald Group Publishing, 2008.
- J.D.M. Rennie. Jacobian of the singular value decomposition with application to the trace norm distribution. February 2006. URL <http://qwone.com/~jason/writing/svdJacobian.pdf>.
- K. Rogoff. The purchasing power parity puzzle. *Journal of Economic Literature*, **34**: 647–668, 1996.
- X. Sala-I-Martin, G. Doppelhoffer, and R.I. Miller. Determinants of long-term growth: A Bayesian averaging of classical estimates (BACE) approach. *American Economic Review*, 94(4):813–835, September 2004. doi: 10.1257/0002828042002570.
- P.C. Schotman and H.K. Van Dijk. A Bayesian analysis of the unit root in real exchange rates. *Journal of Econometrics*, **49**:195–238, 1991.
- C.A. Sims. Macroeconomics and reality. *Econometrica*, 48(1):1–48, January 1980.
- C.A. Sims. Bayesian skepticism on unit root econometrics. *Journal of Economic Dynamics and Control*, **12**:463–474, 1988.
- C.A. Sims and H. Uhlig. Understanding unit rooters: a helicopter tour. *Econometrica*, **59**:1591–1599, 1991.
- J.H. Stock and M. Watson. Combination forecasts of output growth in a seven-country data set. *Journal of Forecasting*, 23:405–430, 2004.
- R. Strachan and H.K. Van Dijk. Bayesian averaging over many dynamic model structures with evidence on the Great Ratios and liquidity trap risk. Tinbergen Institute report 2008-096/4, Erasmus University Rotterdam, 2008.
- R.W. Strachan and H.K. Van Dijk. Bayesian model selection with an uninformative prior. *Oxford Bulletin of Economics and Statistics*, 65(s1):863–876, 2003.
- R.W. Strachan and H.K. Van Dijk. Evidence on features of a DSGE business cycle model from Bayesian model averaging. *International Economic Review*, 54(1):385–402, 02 2013.
- M.A. Tanner and W.H. Wong. The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association*, **82**:528–550, 1987.
- N. Terui and H.K. Van Dijk. Predictability in the shape of the term structure of interest rates. *International Journal of Forecasting*, **18**:421–438, 2002.
- A. Timmermann. Forecast combinations. In G. Elliot, C.W.J. Granger, and A. Timmermann, editors, *Handbook of Economic Forecasting*. North-Holland, 2006.

- H.K. Van Dijk and T. Kloek. Further experience in Bayesian analysis using Monte Carlo integration. *Journal of Econometrics*, 14(3):307–328, 1980.
- M. Villani. Bayesian reference analysis of cointegration. *Econometric Theory*, 21(02): 326–357, 2005.
- M. Villani. Steady-state priors for vector autoregressions. *Journal of Applied Econometrics*, 24(4):630–650, 2009. doi: 10.1002/jae.1065.
- M. West. Bayesian inference in cyclical component dynamic linear models. *Journal of the American Statistical Association*, **90**:1301–1312, 1995.
- M. West and J. Harrison. *Bayesian forecasting and dynamic models*, 2 edn. Springer-Verlag, New York, 1997.
- A. Zellner. *An Introduction to Bayesian Inference in Econometrics*. John Wiley and Sons, New York, 1971. reprinted in 1996.
- A. Zellner, L. Bauwens, and H.K. Van Dijk. Bayesian specification analysis and estimation of simultaneous equation models using Monte Carlo methods. *Journal of Econometrics*, 38(1):39–72, 1988.
- A. Zellner, T. Ando, N. Baştürk, L. Hoogerheide, and H.K. Van Dijk. Bayesian analysis of instrumental variable models: Acceptance-rejection within Direct Monte Carlo. *Econometric Reviews*, 33:3–35, 2014.
- E. Zivot. A Bayesian analysis of the unit root hypothesis within an unobserved components model. *Econometric Theory*, **10**:552–578, August 1994.

The Tinbergen Institute is the Institute for Economic Research, which was founded in 1987 by the Faculties of Economics and Econometrics of the Erasmus University Rotterdam, University of Amsterdam and VU University Amsterdam. The Institute is named after the late Professor Jan Tinbergen, Dutch Nobel Prize laureate in economics in 1969. The Tinbergen Institute is located in Amsterdam and Rotterdam. The following books recently appeared in the Tinbergen Institute Research Series:

- 593 Y. YANG, *Laboratory Tests of Theories of Strategic Interaction*
- 594 M.P. WOJTOWICZ, *Pricing Credits Derivatives and Credit Securitization*
- 595 R.S. SAYAG, *Communication and Learning in Decision Making*
- 596 S.L. BLAUW, *Well-to-do or doing well? Empirical studies of wellbeing and development*
- 597 T.A. MAKAREWICZ, *Learning to Forecast: Genetic Algorithms and Experiments*
- 598 P. ROBALO, *Understanding Political Behavior: Essays in Experimental Political Economy*
- 599 R. ZOUTENBIER, *Work Motivation and Incentives in the Public Sector*
- 600 M.B.W. KOBUS, *Economic Studies on Public Facility use*
- 601 R.J.D. POTTER VAN LOON, *Modeling non-standard financial decision making*
- 602 G. MESTERS, *Essays on Nonlinear Panel Time Series Models*
- 603 S. GUBINS, *Information Technologies and Travel*
- 604 D. KOPÁNYI, *Bounded Rationality and Learning in Market Competition*
- 605 N. MARTYNOVA, *Incentives and Regulation in Banking*
- 606 D. KARSTANJE, *Unraveling Dimensions: Commodity Futures Curves and Equity Liquidity*
- 607 T.C.A.P. GOSENS, *The Value of Recreational Areas in Urban Regions*
- 608 L.M. MARĆ, *The Impact of Aid on Total Government Expenditures*
- 609 C. LI, *Hitchhiking on the Road of Decision Making under Uncertainty*
- 610 L. ROSENDAHL HUBER, *Entrepreneurship, Teams and Sustainability: a Series of Field Experiments*
- 611 X. YANG, *Essays on High Frequency Financial Econometrics*
- 612 A.H. VAN DER WEIJDE, *The Industrial Organization of Transport Markets: Modeling pricing, Investment and Regulation in Rail and Road Networks*
- 613 H.E. SILVA MONTALVA, *Airport Pricing Policies: Airline Conduct, Price Discrimination, Dynamic Congestion and Network Effects.*
- 614 C. DIETZ, *Hierarchies, Communication and Restricted Cooperation in Cooperative Games*
- 615 M.A. ZOICAN, *Financial System Architecture and Intermediation Quality*
- 616 G. ZHU, *Three Essays in Empirical Corporate Finance*
- 617 M. PLEUS, *Implementations of Tests on the Exogeneity of Selected Variables and their Performance in Practice*

- 618 B. VAN LEEUWEN, *Cooperation, Networks and Emotions: Three Essays in Behavioral Economics*
- 619 A.G. KOPÁNYI-PEUKER, *Endogeneity Matters: Essays on Cooperation and Coordination*
- 620 X. WANG, *Time Varying Risk Premium and Limited Participation in Financial Markets*
- 621 L.A. GORNICKA, *Regulating Financial Markets: Costs and Trade-offs*
- 622 A. KAMM, *Political Actors playing games: Theory and Experiments*
- 623 S. VAN DEN HAUWE, *Topics in Applied Macroeconometrics*
- 624 F.U. BRÄUNING, *Interbank Lending Relationships, Financial Crises and Monetary Policy*
- 625 J.J. DE VRIES, *Estimation of Alonso's Theory of Movements for Commuting*
- 626 M. POPLAWSKA, *Essays on Insurance and Health Economics*
- 627 X. CAI, *Essays in Labor and Product Market Search*
- 628 L. ZHAO, *Making Real Options Credible: Incomplete Markets, Dynamics, and Model Ambiguity*
- 629 K. BEL, *Multivariate Extensions to Discrete Choice Modeling*
- 630 Y. ZENG, *Topics in Trans-boundary River sharing Problems and Economic Theory*
- 631 M.G. WEBER, *Behavioral Economics and the Public Sector*
- 632 E. CZIBOR, *Heterogeneity in Response to Incentives: Evidence from Field Data*
- 633 A. JUODIS, *Essays in Panel Data Modelling*
- 634 F. ZHOU, *Essays on Mismeasurement and Misallocation on Transition Economies*
- 635 P. MULLER, *Labor Market Policies and Job Search*
- 636 N. KETEL, *Empirical Studies in Labor and Education Economics*
- 637 T.E. YENILMEZ, *Three Essays in International Trade and Development*
- 638 L.P. DE BRUIJN, *Essays on Forecasting and Latent Values*
- 639 S. VRIEND, *Profiling, Auditing and Public Policy: Applications in Labor and Health Economics*
- 640 M.L. ERGUN, *Fat Tails in Financial Markets*
- 641 T. HOMAR, *Intervention in Systemic Banking Crises*
- 642 R.L. LIT, *Time Varying Parameter Models for Discrete Valued Time Series*